DOCUMENT RESUME

ED 394 971 SP 036 687

TITLE	An Independent Evaluation of the Kentucky Instructional Results Information System (KIRIS).
INSTITUTION	Western Michigan Univ., Kalamazoo. Evaluation Center.
SPONS AGENCY	Kentucky Inst. for Education Research, Frankfort.
PUB DATE	Jan 95
NOTE	99p.; For related documents, see SP 036 685-694.
PUB TYPE	Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC04 Plus Postage.
DESCRIPTORS	*Educational Assessment; Elementary Secondary Education; *Evaluation Utilization; Information Systems; *Outcomes of Education; Performance Based Assessment; *Program Evaluation; Rewards; Sanctions; State Departments of Education; State Standards; *Student Evaluation; Test Reliability; Test Validity
IDENTIFIERS	Kentucky; *Kentucky Education Reform Act 1990; *Kentucky Instructional Results Information System

ABSTRACT

This document contains the executive summary and the detailed report which provide an independent evaluation of Kentucky's new system for assessing student performance, the Kentucky Instructional Results Information System (KIRIS). The summary gauges progress to date, highlights some strengths to be built on and problems to be solved, and provides suggestions for improvement. The main report delineates the information needed to conduct a comprehensive evaluation of the reliability and validity of KIRIS. The evaluation was accomplished through study of technical documentation, reports, newspaper articles, and monographs; observation of KIRIS performance events exercises; interviews with key persons involved in development and implementation; observation of Kentucky educators meetings about KIRIS; and informal discussions with and attitudinal surveys from teachers, parents, principals, superintendents, and district assessment coordinators. In general, it was found that KIRIS was consistent with the requirements set forth by the Kentucky Education Reform Act (KERA) and that most stakeholders had some understanding of the rewards and sanctions component. Although Kentucky educators had been involved in the design and development of KIRIS, some teachers believed that questions on the assessment were written by persons with little or no knowledge of Kentucky. A need was found for much better organization and improved balance of the information generated. Teacher time spent on the assessment was deemed useful and reasonable, but the accountability index was found to be slow in providing teachers with timely feedback. Recommendations are offered about steps that could be taken to address the continuing needs for improvement. Appendices include a listing of the evaluation procedures and sources of evidence, a graphic outline of the system, and a letter responding to a citizen's concern about a particular question on the assessment. (NAV)

ED 394 971

THE KENTUCKY INSTITUTE FOR EDUCATION RESEARCH

An Independent Evaluation of the Kentucky Instructional Results Information System (KIRIS)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

R. Vanderatz

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

conducted by
The Evaluation Center
Western Michigan University

for the
Kentucky Institute for Education Research
146 Consumer Lane, Frankfort, Kentucky 40601

January 1995

SE036687

BEST COPY AVAILABLE

BOARD OF DIRECTORS

Chair
Ben Richmond
Urban League of Louisville
1535 West Broadway
Louisville, KY 40203-3516

Vice Chair
Gary Dodd
CM Management Services
698 Perimeter Drive, Suite 200
Lexington, KY 40517

Secretary
Robert F. Sexton
The Prichard Committee
for Academic Excellence
P. O. Box 1658
Lexington, KY 40592

Treasurer
Doug Kuelpman
United Parcel Service
1400 North Hurstbourne Parkway
Louisville, KY 40223

Lila Bellando
Churchill Weavers
P. O. Box 30
Berea, KY 40403

Barbara Deeb
WKYU-TV, 1 Big Red Way
Bowling Green, KY 42101

Jane Joplin Evans
515 North Main Street
Somerset, KY 42501

Judy B. Thomas
Ashland Oil Foundation
P. O. Box 391
Ashland, KY 41141

Fred D. Williams
70 Pentland Place
Ft. Thomas, KY 41075

Amy Helm Wilson
Murray Ledger & Times
1001 Whitnell Avenue
Murray, KY 42071

Joe Wright
Star Route
Harned, KY 40144

Executive Director
Roger S. Pankratz, Ph.D.
KY Institute for Education Research
146 Consumer Lane
Frankfort, KY 40601

THE KENTUCKY INSTITUTE FOR EDUCATION RESEARCH

146 Consumer Lane • Frankfort, Kentucky 40601 • 502-227-9014 • Fax 502-227-8976

THE
KENTUCKY
INSTITUTE
FOR
EDUCATION
RESEARCH

146 Consumer Lane, Frankfort, KY 40601- 502/227-9014 - FAX 502/227-8976

PREFACE

This independent evaluation of the Kentucky Instructional Results Information System (KIRIS) was produced by the Kentucky Institute for Education Research in partial fulfillment of its mission by Executive Order to "conduct an in-depth evaluation of the impact of the Kentucky Education Reform Act (KERA) on students, individual schools, school systems and educators . . ." The evaluation study was conducted by The Evaluation Center at Western Michigan University for the Kentucky Institute for Education Research. The Institute is pleased to provide this report to citizens, elected leaders and educators as the expert judgement of an evaluation team about the progress of the development and implementation of KIRIS through December 1994 and how Kentucky's new school assessment and accountability system can be strengthened and improved. The Kentucky Institute for Education Research endorses the contents of this report as the work of dedicated and expert professionals and deserves thoughtful consideration by policy makers, educators and the public.

EXECUTIVE SUMMARY

An Independent Evaluation of the Kentucky Instructional Results Information System (KIRIS)

Report Submitted
by
The Evaluation Center
Western Michigan University
to
The Kentucky Institute for Education Research
January 1995

Purpose

This executive summary and the detailed report on which it is based provide an evaluation of Kentucky's new system for assessing student performance, the Kentucky Instructional Results Information System (KIRIS). The reports are intended to provide useful feedback to parents, students, teachers, principals, state officials, and other Kentucky education policymakers at an early stage in the development of the new education assessment system. This summary gauges progress so far, highlights some strengths to be built upon and problems to be resolved, and provides some suggestions for improvement. Our main report delineates the information needed to conduct a comprehensive evaluation of the reliability and validity of KIRIS.

KIRIS

KIRIS is Kentucky's legislatively mandated effort to develop a state-of-the-art, "high stakes" student assessment system that is primarily performance based--that is, one that uses a variety of students' performances of tasks, instead of multiple-choice tests, to assess student learning. This new system is used to drive curriculum, instruction, and school administration to ensure that all schools meet the "goals for the Commonwealth's schools" (KRS 158.6451).

Through KIRIS, the Commonwealth (1) provides an annual assessment of the performance of Kentucky students at selected grade levels, (2) holds each school accountable for achieving the reform goals, (3) administers economic rewards and sanctions based on the test data and noncognitive information, and (4) promotes and supports the use of performance assessment as an integral part of classroom instruction. The Kentucky Education Reform Act requires that KIRIS be used to grant economic rewards to schools that showed improvement over a threshold level and to deliver state assistance and sanctions to schools that do not reach their threshold level. For the 1992-1994 biennium, schools with growth rates below expectations will receive planning grants and a "distinguished educator" to help them improve. The rewards and sanctions make the Kentucky educational reform a "high stakes" program and KIRIS a "high stakes" assessment.

The Accountability Index. A school's accomplishment is described by an accountability number, which is a composite of six equally-weighted component scores. In the initial year of the program, 1991-1992, five of the components were cognitive: reading, mathematics, social studies, science, and writing. There was one noncognitive component, which itself is a composite of attendance, retention, dropout rate, and transition to adult life. This noncognitive component counts as 1/6 of the total score on the accountability index. The noncognitive component and the calculation of the school improvement goals lie outside of KIRIS itself.

Difficulty of Implementing a Totally New and Innovative Performance-Based Assessment System

We recognize the long and hard work put into the KIRIS assessment system by the Kentucky Department of Education (KDE) and its outside assessment contractor, Advanced Systems in Measurement and Evaluation (ASME). The use of performance assessments in large-scale assessment systems in the U.S. is relatively new compared to the more traditional multiple-choice tests. Neither education and testing agencies nor the measurement profession has solved the many technical and operational problems with large-scale use of performance-based assessments. KDE and ASME might have preferred to proceed slowly when implementing the new performance assessment system. However, in the face of the legislative mandate and the press for reform in Kentucky, KDE and ASME postponed much of the needed research and development of assessment questions and implemented the legislatively mandated performance-based system at a very fast pace. KDE continues to work on the needed assessment research and development.

KDE and its collaborators have exerted herculean efforts and have accomplished much. They have developed a complex system of performance assessment and applied it on a statewide basis. They have encountered problems, which is to be expected in so massive and fast-paced an undertaking on the cutting edge of technology. KDE is and must be in a constant state of innovation, trial and error testing, and refinement of the measurement system.

KDE and ASME have only had since 1991 to design, develop, implement, explain, and obtain stakeholder acceptance of the concept of the new performance assessment system. During that time they had to train the participants and work through the inevitable problems of developing, administering, scoring, and reporting performance-based assessments.

KDE had to (1) work out the logistics concerning the new assessment (open-ended questions, performance events, portfolios); (2) train teachers to prepare students to submit portfolios; (3) develop a system to train teachers to evaluate portfolios; (4) set standards without a history of pertinent data; (5) define successful students; (6) develop guidelines for the participation of special education students; (7) develop the weights for the accountability index; (8) establish thresholds; (9) develop a rewards system; (10) develop the criteria for sanctions; and (11) deal with other issues mandated by the legislature. It would not be appropriate for any party to use this report to destroy the valuable progress that KDE is making in developing KIRIS into an educationally sound system of assessment and accountability.

The Need for Evaluation of KIRIS

In view of the prodigious amount of work under way, the innovativeness of the effort, and the high stakes involved for Kentucky schools, it was important that KIRIS be subjected to independent evaluations. Since the system was being implemented while much of the supporting research and development was under way, it was important to get feedback on strengths, weaknesses, and issues requiring attention. The nonpartisan Kentucky Institute for Education Research (KIER), with endorsement by the State Board for Elementary and Secondary Education, commissioned this study.

The Evaluators

KIER contracted with the Western Michigan University Evaluation Center to conduct this evaluation. The evaluation team included Mark Fenster (political scientist from Western Michigan University), Anthony Nitko (University of Pittsburgh and past president of the National Council on Measurement in Education), Daniel Stufflebeam (Western Michigan University and past chair of the national Joint Committee on Standards for Educational Evaluation), and William Wiersma (educational researcher from the University of Toledo). The evaluation team thanks Robert Meyer (University of Chicago) for providing expert consulting advice. No member of the team was involved in the development of KIRIS, and none has any vested stake in the Kentucky education reforms. All, however, are dedicated educators who very much want to see the Kentucky educational reforms succeed to the benefit of all the children in the public schools of the Commonwealth.

The Evaluation Procedure

During the period of May through December 1994, our team evaluated KIRIS by using the following procedures:

1. Collected and reviewed a wide range of relevant technical documentation of KIRIS
2. Collected and reviewed reports, newspaper articles, monographs, etc., having to do with KIRIS
3. Observed an administration of KIRIS performance events exercise
4. Interviewed key persons involved in the development of KERA and KIRIS
5. Interviewed key persons involved in the implementation and use of KIRIS
6. Observed meetings of Kentucky educators concerned with KIRIS
7. Conducted informal discussions with several groups, including teachers, parents, principals, superintendents, and district assessment coordinators

8. Surveyed selected groups--legislators, parents, district assessment coordinators, and superintendents--to obtain their views on KIRIS

9. Developed draft reports and subjected them to review by experts on education in Kentucky

Reviewers

Prior drafts of this report were reviewed, under agreement to keep the report confidential, by Ken Draut (Henry County Schools), Ray Nystrand (University of Louisville), Robert Rodosky (Jefferson County Public Schools), Skip Kiefer (University of Kentucky), Roger Pankratz (KIER), James Craig (Western Kentucky University), Edward Reidy (Kentucky Department of Education), Neal Kingston (Kentucky Department of Education), and Brian Gong (Kentucky Department of Education). We thank the reviewers for providing valuable feedback. We have attempted to consider and, as we deemed appropriate, to address all their concerns. However, the Western Michigan University Evaluation Center and the evaluation team named above are responsible for the contents of the report and bear sole responsibility for any factual errors or ambiguities.

The Evaluation Topics

The above procedures, which are presented in greater detail in the main report, were used to address the contracted evaluation questions. KIER and the evaluation team jointly determined that this evaluation should address questions concerned with the following:

- Consistency with Legislative Mandate
- Understanding and Confidence of Stakeholders in KIRIS Assessment
- Involvement of Teachers and Principals in Design and Development of KIRIS
- Accuracy, Accessibility, and Clarity of Documentation
- Impact of KIRIS Accountability Policies on Students, Teachers, and Schools
- Technical Adequacy of the KIRIS Assessment

The full report responds directly to specific questions in each of these categories. Here we summarize only the main findings.

Limitations of Our Study

This executive summary provides only a snapshot of the contents of the main report.

Also, our evaluation has limitations that the reader should keep in mind. We conducted the evaluation at a point in the development of KIRIS when much of the needed technical

information was not yet available. This evaluation covers information gathered on KIRIS through December 1994. Much of the technical data required to evaluate the validity and reliability of the KIRIS assessments will not be available until the Biennium I Technical Report is released. For this reason, the evaluation mainly reports on the first two years of the KIRIS assessment (1991-1992, 1992-1993). We had quite limited time and resources to devote to the evaluation. While we obtained valuable input from many Kentucky educators and other stakeholders, their views are not necessarily representative of those of stakeholders at large in Kentucky. In the final analysis, our reports contain our best judgments and recommendations based on what we were able to learn through an intense effort to understand and assess KIRIS based on our reviews of available documentation, our surveys of groups of key stakeholders, our conduct of focus groups and interviews, and our deliberations as a team.

Our Main Findings

For each of the study topics, we list points of both strengths and weaknesses. In the ensuing section we offer our ideas about what steps could be taken to strengthen the assessment program.

Consistency with Legislative Mandate.

1. Overall, KIRIS is consistent with the Kentucky Educational Reform Act.

   1.1 On the major issue of performance-based, high stakes assessment, the Kentucky Department of Education has pursued the intent of the legislation. The Department was required to produce a fundamentally different kind of assessment for Kentucky students than the previously used state assessment tests. With KIRIS assessment, the Department of Education produced an assessment broadly consistent with legislative mandates.

   1.2 The legislation stipulated that the assessments were to provide the state with national comparisons similar to those provided by the National Assessment of Educational Progress (NAEP--a federal assessment program providing benchmark information on student achievement). KDE provided national comparisons for two subject-matter areas in the 1992-1993 technical report. We understand that KDE plans to issue additional comparisons of KIRIS results with NAEP results when future NAEP results become available.

Understanding and Confidence of Stakeholders in the KIRIS Assessment.

2. Most of the people who provided data for our study have some understanding of the rewards and sanctions component of the KIRIS assessment. However, specifics regarding rewards and sanctions are probably known only to a limited number of people (Department of Education personnel, superintendents and district

assessment coordinators in some districts, some teachers, some principals, state legislators sitting on accountability committees, and some testing experts).

3. All the reviewed evidence suggests that principals, coordinators, superintendents, teachers, school council parents, public school parents, legislators, and the general public have serious questions concerning the legitimacy, validity, reliability, fairness, and usefulness of the KIRIS assessment. The groups surveyed perceived student performance on the KIRIS assessment as the measure **least** likely to provide a reliable indicator of student learning, compared to other commonly available indicators such as high school completion rate. The KDE will need to convince Kentucky educators that KIRIS is a sound basis for judging school effectiveness if this system is to become a valued part of the education reform process.

Involvement of Teachers and Principals in Design and Development of KIRIS.

4. As described in the 1991-1992 Technical Report, advisory committees were established for reading, mathematics, science, and social studies. Representatives of KDE, teachers, curriculum coordinators, and Kentucky Education Association members sat on the committees. KDE added additional committees when other subjects were added to the assessment.

5. However, some of the teachers we communicated with were unaware of the input other Kentucky teachers had through these committees. Despite the committee system and the input of Kentucky educators into the review process of the KIRIS assessment, some teachers perceived that questions on the assessment were constructed by outsiders with little or no knowledge of Kentucky. Clearly, the perception of some teachers is at odds with the fact of educator involvement in KIRIS. This underscores the importance of continuing to involve and inform teachers and other educators in the ongoing process of assessment development.

Accuracy, Accessibility, and Clarity of Documentation.

6. KDE has developed substantial technical information about KIRIS, given the early stage of development. As the program develops, there will be a continuing and growing need for technical information. We have outlined our view of what will be needed in our full report.

7. Also, there is a need for much better organization and improved balance of the information. While there is a considerable amount of technical data on the KIRIS assessment available in various places, it is difficult for anyone reviewing the program to compile all the relevant information. The technical reports do not provide a complete perspective on the weaknesses as well as the strengths of the KIRIS assessment results and on the accountability index.

Impact of the KIRIS Accountability Policies on Students, Teachers, and Schools.

8. Students experienced more writing and group work under the reforms. Teachers, district assessment coordinators, and superintendents report almost unanimously that writing has improved, and the writing improvement was over and above what would have been expected of most school children of the same age.

9. Portfolios of students' written work have great instructional potential. However, portfolio scores vary considerably depending on which teacher is scoring the portfolio, making these scores less reliable than other forms of assessment.

10. The time and effort KDE invests in training teachers and that teachers spend marking the KIRIS portfolios, in our judgment, is probably useful as inservice education for teachers. We judge that the amount of instructional time teachers spend on the KIRIS assessment is reasonable.

11. The accountability index is influenced by factors beyond a school's control, but these are not taken into account when the index is interpreted. (Perhaps this is because the legislation does not require these factors to be taken into account.) Among the factors not considered are adequacy of resources, changes in the economic climate of a community, and changes in student mobility. However, the state maintains a mechanism by which a school's authorities can appeal such matters. That is, if a school believes the state's finding that the school failed to achieve the goal set for it by the state is due to factors beyond the control of the school, it can appeal the state's determination.

12. The accountability index does not provide teachers with timely feedback that is directly usable for improving classroom activities. While the index is not designed to provide such feedback, many of the educators with whom we communicated want more such feedback than the accountability index and the other aspects of KIRIS provide.

13. There is disagreement on the question of whether the system of rewards and sanctions will help improve the quality of education in Kentucky. District assessment coordinators think that rewards and sanctions will help improve education. Results for superintendents vary by survey. Teachers surveyed by KIER say the rewards and sanctions will **not** help improve education.

14. There is concern but as yet limited evidence about whether the administration of rewards and sanctions is fair to schools with large numbers of economically disadvantaged students, high turnover rates, or a very small number of students. We understand that KDE plans to provide further information on this important question in the near future.

15. The legislative intent of integrating assessment and feedback into the instructional process at every grade level has not been achieved. Teachers need more assistance than the Department of Education has so far been able to provide to embed performance assessments into the instructional process as was envisaged in the legislation.

Technical Adequacy of the KIRIS Assessment.

16. On the whole we judged the KIRIS assessment tasks to be technically well crafted (the questions are clear and appropriate for the age group, the scoring rules are valid, and instructions are easy for students to follow).

17. The open-ended questions (those requiring a written answer) generally meet currently accepted technical item-writing standards for open-ended questions.

18. The district assessment coordinators and the superintendents overwhelmingly prefer a longitudinal approach (tracking the same group of individual students as they progress through the grades) over a cohort approach (comparing each group of 4th graders to those of previous years) for assessing a school's growth or change. In the opinion of the research team, longitudinal analysis gives a better picture of what impact the school is having on a group of students as it goes through the educational system, although it is more difficult and costly to implement. We believe that effective use of the longitudinal approach would require that assessments be administered at least to students at every other grade level and preferably at every grade level. It may also entail developing a growth scale on which a school's progress may be assessed. We note here that KDE made a deliberate policy decision early in the reform movement not to use the longitudinal model to evaluate growth in assessment scores.

19. The 1993-1994 KIRIS assessment included several modes: extended answer open-ended tasks, shorter answer open-ended tasks, multiple-choice tasks, portfolios, and performance events. This diversity of approaches is a strength of the scheme. The multiple modes of assessment approach is supported by educational assessment specialists because it enhances the validity of the results. Validity is enhanced by allowing students opportunities to demonstrate their abilities in a variety of ways over an appropriate range of knowledge and skills. The proposed 1994-1995 KIRIS assessment will allow students fewer opportunities to demonstrate their abilities compared to the 1993-1994 KIRIS assessment due to the elimination of multiple-choice items. We think it was a mistake that KDE did not count the performances on multiple-choice items in computing the school accountability index; we think it would be a further mistake if KDE were to eliminate the multiple-choice items altogether. In addition, plans to increase the number of short-answer questions, instead of increasing the more in-depth performance components, narrows the modes of the assessment. The general point is that it will be desirable to broaden the assessment modes used.

20. The reliability of the accountability index is problematic for us. KDE has reported impressive reliabilities that reach or exceed .90, a level generally considered to be acceptable for use in high stakes decisions. However, because of the particular statistical model employed, there are unresolved questions about whether the high reliability estimates are indicative of the actual reliability. These concern, for example, whether to treat items or students as fixed, how agreements among raters are taken into account, and whether student scores should be estimated with regression.

21. Setting aside the issue of the statistical model for estimating reliability, it is clear that taken by themselves two of the three components of the KIRIS accountability index are not sufficiently reliable to be used in a high stakes assessment. These two components are the writing portfolio and performance events. We question whether the combination of these two components and the open-ended questions, which do evidence good reliability, give the Commonwealth a sufficiently reliable index for administering rewards and sanctions to schools. More reliability evidence is needed on this matter If the index is unreliable, then its validity is open to question since validity depends in part on reliability. The issues of the validity as well as the stability of the index require careful study, so that all stakeholders can be reassured that it provides a credible basis for administering rewards and sanctions or so that it can be corrected as needed.

Main Recommendations

The preceding assessment of strengths and weaknesses denotes that efforts to improve the KIRIS need to be continued if it is to provide a defensible basis for high stakes decisions and if it is to contribute productively to improving classroom instruction. In this section we offer our ideas about some of the steps that could be taken to address the continuing needs for improvement. While we have not had the time and resources to thoroughly develop these recommendations and to compare them to other possible improvement steps, we offer them in the spirit of helping Kentucky stakeholders to consider how best to continue improving KIRIS.

Additional Information and/or Reporting is Needed.

1. There is a need to evaluate and address as appropriate concerns about the use of the accountability system. Among the concerns heard in our exchanges with Kentucky educators are that the current KIRIS assessment

   - narrows the curriculum
   - produces undue stress, especially on 4th grade teachers
   - yields an unstable index and unfair basis for accountability in those schools where individual student populations may vary widely from year-to-year and grade-to-grade
   - does not provide parents with reliable individual level student scores

2. The Commonwealth should investigate and report whether inner-city urban schools are being unfairly sanctioned because they have a more difficult educational task than the more stable schools. We understand that KDE plans to undertake such investigation following the completion of the first accountability cycle. However, this does not mitigate the fact that KIRIS results are being used in high stakes decision making before the needed evidence on the validity of KIRIS for this purpose could be obtained.

3. An index should be developed to report on the progress of students in meeting each of the four reform goals. It would also be desirable to report performance of schools on clusters of academic expectations.

4. Document and fully publicize the degree of interpretive and consequential validity of KIRIS. Also, document its instructional utility. Publicized reports should explain the appropriate cautions in using KIRIS results to claim educational improvement in Kentucky.

5. Continue to develop methods for reporting to schools on how they could use the KIRIS results to alter teaching and to improve student learning.

Training of Stakeholders.

6. Given the weight of the writing portfolio in the accountability index, we recommend that the state continue to place great importance on the training of teachers to understand the deeper meaning of student writing and to score the writing portfolio.

7. Because of the instructional value of portfolios and the importance of having teachers seriously evaluate the best work of their students, teachers in Kentucky should continue to score the portfolios, even though scores of the same portfolios may vary from one teacher to the next and are, therefore, less reliable.

8. Expand on the steps being taken to involve and inform Kentucky educators about issues and developments in KIRIS. As much as possible, bring them into the partnership for developing and using a sound accountability index and helping to communicate KIRIS results to parents and other interested groups. Involve all Kentucky teachers in the process of crafting tasks that will be used in the operational assessment instruments.

9. Expand activities to help Kentucky teachers to incorporate the performance tasks and higher quality continuous assessments into their regular classroom instruction for all grade levels, as envisioned by the KERA.

Technical Issues.

10. The technical reports should be organized so that an outside technical reader can evaluate the reliability and validity of the KIRIS results for achieving the uses and interpretations claimed for them. They also should summarize all the research results underpinning the program. There should be sections in the technical reports that point out problems and inconsistencies with the assessment. In general, they should include all the relevant technical information specified in the Standards for Educational and Psychological Testing (1985).

11. Beyond the requirements of the current standards, we suggest that KDE calculate and report reliability estimates for the accountability index based on a model that considers both students and items to be random sources of error, along with the estimates they now report using a model that considers students and items to be fixed factors. While this dual reporting would not resolve the debate about which model is the more appropriate, it would show readers the consequences to the reliability estimate of using one model or the other. We continue to believe that students and items should be considered as random sources of error in the generalizability model employed, since scores from one set of students and one set of assessment items are obtained in one year to set a threshold for evaluating the performance of a different set of students on a different set of assessment items in a subsequent year.

12. Continue to use the performance events. If the necessary approval can be obtained, we think it would be desirable to use the performance events to assess individual abilities to work collaboratively in groups as well as perform important learning targets as derived from the Kentucky goals and academic expectations. We note that KDE's current practice of not assessing students' ability to work effectively in groups is consistent with what the legislation permits.

13. Increase the priority and human energy resources devoted to analyzing data that support the technical underpinnings of the assessment results. This may or may not require an expanded staff. (We perceive that this change may already be under way, e.g., through studies relating the KIRIS results to American College Test [ACT] scores and through the Office for Education Accountability's study of the assessment.)

14. Provide evidence to demonstrate that the accountability index has a level of validity sufficient for use in high stakes decisions such as those affecting rewards and added resources such as planning grants and assignment of distinguished educators. Alternatively, if the necessary level of validity is not attained, do not continue to use the index for such decisions and actions until it is improved.

15. A key step toward improving validity will be to obtain external confirmation, as, for example, from the ACT, that the accountability index does manifest an

acceptable level of reliability. Reliability is a necessary, but not sufficient, condition for validity. We recommend that KDE consult with a nationally recognized psychometrician who specializes in generalizability theory. The specialist should evaluate the statistical model, the estimated score procedures, and the design of the generalizability studies.

16. We think the decision not to include enhanced multiple-choice items in the index also limits the validity that could be attained, e.g., through improving both content coverage and reliability. We recommend, therefore, that KDE reassess the decision not to use enhanced multiple-choice test items, along with the short answer and performance assessments, in assessing student progress and computing the accountability index.

17. In the spirit of KERA's concern for authentic assessment, we also recommend that KDE at least consider increasing the emphasis on performance assessments that require speaking, developing products, organizing and planning activities, etc., compared to the heavy emphasis now given to performance assessments that require only written responses. We acknowledge that KDE and its contractor would need to conduct relevant research and development to fulfill this recommendation.

18. Consider using a longitudinal model to assess change in a school's accountability index.

Summation

We are mindful of the Kentucky Department of Education's important responsibility of informing Kentucky citizens about the outcomes of the Kentucky Education Reform Act. We fully support the goal of developing more effective ways to inform the public, educational policymakers, and educators about the progress of Kentucky students. We commend the state legislature, the state school board, the Department of Education, and the KIRIS contractors, Advanced Systems for Measurement and Evaluation (ASME), for starting this bold innovation aimed at breaking the mold of using only multiple-choice tests to assess school outcomes and installing instead an open-ended and performance-based assessment system aimed at furthering educational reform. We think that much of the new assessment system is conceptually consistent with the aims of the legislation but that many critical problems and issues must be effectively addressed. We endorse the Department's continuing efforts to assure that KIRIS will fulfill its important role of driving curriculum and classroom instruction so that all students in Kentucky will meet the state's standards for educational achievement.

The Kentucky Department of Education and ASME have achieved much and are to be commended. They need more time to resolve a range of difficult technical, utility, and communication issues in KIRIS. We hope this report will be of use in that process.

AN INDEPENDENT EVALUATION OF THE KENTUCKY INSTRUCTIONAL RESULTS INFORMATION SYSTEM (KIRIS)

Conducted for

The Kentucky Institute for Education Research
Frankfort, KY

by

The Evaluation Center
Western Michigan University

January 1995

Table of Contents

Preamble 1

Context of the Evaluation 1

The Need for Evaluation of KIRIS 2

The Evaluation Team 2

Reviewers 3

Purpose of This Report 3

Audience 3

Background 3

- Description of the KIRIS Assessment 5
- How the Assessment Tasks Within a Cognitive Measure are Weighted 7
- How the Components Within the Noncognitive Measures are Weighted 8
- Recent Changes in the Assessment 9

Evaluation Questions 10

- Evaluation Procedures and Sources of Evidence 11

Limitations of Our Study 12

Answers to Evaluation Questions 12

- Understanding and Confidence of Stakeholders in the KIRIS Assessment 16
- Involvement of Stakeholders in the Design and Development of KIRIS 21
- Accuracy and Accessibility of Documentation 21
- Impact of KIRIS Accountability Policies 24
- Technical Adequacy of the KIRIS Assessment 36

Summary and Conclusions 66

Preamble

This is an external evaluation of the progress made by the Kentucky Department of Education (KDE) since 1991 in developing and using the Kentucky Instructional Results Information System (KIRIS). The Evaluation Center (EC) at Western Michigan University prepared the report pursuant to its contract with the Kentucky Institute for Education Research (KIER). The Center distributed prior drafts of the evaluation report to selected reviewers to obtain their critical reactions and thereby to attempt to correct any factual errors and to clarify areas of ambiguity. In addition to this report, EC is also providing an executive summary to KIER.

Context of the Evaluation

This report must be considered in the dynamic context of the educational reforms under way in Kentucky since 1990. The scale of these reforms is massive and unprecedented for any state. In 1989 the Kentucky Supreme Court declared the Commonwealth's existing rules and procedures for financing schools and delivering educational services to be unconstitutional. In 1990 the state legislature passed the Kentucky Education Reform Act (KERA), which mandated a total overhaul of the K-12 public education system and was designed to result in equitable educational services to all students.

The KDE is leading the effort to reform the Commonwealth's K-12 education system through a fast-paced, highly financed, and labor intensive process of educational change. The main features of the new system are (1) prescribed statewide academic expectations; (2) use of a model curriculum framework; (3) a commitment to helping all children to become proficient in performing rigorous state standards that emphasize application of what is learned; (4) heavier concentration of learning resources on students who are not learning up to their potential; (5) extensive parental involvement; (6) site-based management of schools; (7) use of financial incentives to reward staff in schools where student gains are exemplary; and (8) use of sanctions, including the assistance of distinguished educators, to cause staffs in ineffective schools to bring student achievement gains to an acceptable level. The Commonwealth's development of this new education system is important and of national as well as statewide interest, but it is also encountering many difficulties common in development.

KDE has faced difficulties in being required to develop and apply the new education rules and procedures before it can fully field test and validate them. KDE's challenges in this regard have been especially acute in the development of the performance assessment and accountability components of KERA.

The Kentucky Instructional Results Information System (KIRIS) had to comply with legislative mandates, which are evolving; had to provide performance measures, rather than the more readily available and easier to use multiple-choice tests; and had to produce assessments that would be technically defensible and politically credible for making high stakes decisions on rewards and sanctions to schools. This was a huge challenge requiring quick study, outstanding measurement talent, overtime effort by educators throughout the Commonwealth, much money, and patience by those with oversight authority.

Also, KDE had to innovate in developing KIRIS at a time when the field of educational measurement is itself updating the standards for judging education assessment systems (Linn, 1994). The educational measurement profession has recognized that its current Standards for Educational and Psychological Tests (APA, 1985) require review and updating. Because of Kentucky's pioneering work in performance assessment, its experience may possibly contribute to improving the professional standards of the measurement field.

KDE and its collaborators have exerted herculean efforts and have accomplished much. They have developed a complex system of performance assessment and applied it on a statewide basis. They have encountered problems, which is to be expected in so massive and fast-paced an undertaking on the cutting edge of the technology. The KDE is and must be in a constant state of innovation, trial and error testing, and refinement of the measurement system.

This report is intended to give credit to KDE and its collaborators for what has been accomplished in so short a period of time and to take stock of the strengths and weaknesses of the system in its present state of development. Our clear intention is that this report should be used constructively in the ongoing process of the development of KIRIS. It would not be appropriate for any party to use this report to destroy the valuable progress that KDE is making in developing KIRIS into an educationally sound system of assessment and accountability.

The Need for Evaluation of KIRIS

In view of the prodigious amount of work under way, the innovativeness of the effort, and the high stakes involved for Kentucky schools, it was important that KIRIS be subjected to independent evaluations. Since the system was being implemented while much of the supporting research and development were under way, it was important to get feedback on strengths, weaknesses, and issues requiring attention. The nonpartisan Kentucky Institute for Education Research (KIER), with endorsement of the State Board for Elementary and Secondary Education, thus commissioned this study.

The Evaluation Team

The Evaluation Center's evaluation of KIRIS was conducted by a team consisting of Mark Fenster (political scientist from Western Michigan University), Anthony Nitko (University of Pittsburgh and past president of the National Council on Measurement in Education), Daniel Stufflebeam (Western Michigan University and past chair of the national Joint Committee on Standards for Educational Evaluation), and William Wiersma (educational researcher from the University of Toledo). The evaluation team thanks Robert Meyer (University of Chicago) for providing expert consulting advice on the evaluation of KIRIS. No member of the team was involved in the development and the creation of KIRIS, and none has any vested stake in the Kentucky education reforms. All, however, are dedicated educators who very much want to see the Kentucky educational reforms succeed to the benefit of all the children in the public schools of the Commonwealth.

Reviewers

Prior drafts of this report were reviewed, under agreement to keep the report confidential, by Ken Draut (Henry County Schools), Ray Nystrand (University of Louisville), Robert Rodosky (Jefferson County Public Schools), Skip Kiefer (University of Kentucky), Roger Pankratz (KIER), James Craig (Western Kentucky University), Edward Reidy (Kentucky Department of Education), Neal Kingston (Kentucky Department of Education), and Brian Gong (Kentucky Department of Education). We thank the reviewers for providing valuable feedback. We have attempted to consider and, as we deemed appropriate, to address all their concerns. However, the Western Michigan University Evaluation Center and the evaluation team named above are responsible for the contents of the report and bear sole responsibility for any factual errors or ambiguities.

Purpose of This Report

The purpose of this report is to provide KIER with an independent perspective on the merit of KIRIS in its present state of development. We understand that KIER will use this report to help education leaders and school personnel in Kentucky to appreciate and assess what has been accomplished and to bring about needed improvement in KIRIS.

Audience

We hope the report will be useful to the governor, the State Board of Education, the state legislature, the Kentucky Department of Education, teachers, principals, assessment coordinators, superintendents, local school boards, members of the site-based management councils, parents, students, and other stakeholders of the Kentucky education system. The report highlights some noteworthy strengths, issues, and problems we found in KIRIS. We conclude by offering some recommendations to improve KIRIS.

Additionally, educators in general may find the report useful. Performance assessments in the United States have become more frequent in recent years. Kentucky has gone further in performance assessments than most other states in the United States. For this reason, Kentucky is seen as a bellwether for the country to examine the practicality and impact of performance assessments. More broadly, Kentucky has received national and international attention because of its education reform movement and unique approach to assessment and accountability.

Background

The Commonwealth of Kentucky used the Kentucky Essential Skills Test (KEST) in the middle 1980s and the Comprehensive Test of Basic Skills (CTBS-IV) in 1988-1989 and 1989-1990 to assess students. The Commonwealth could take over school districts if their students did not perform satisfactorily on KEST. However, some people in the state thought there were some fundamental problems with an accountability system focused at the district level. Within a district, schools with weak or descending test scores could be counterbalanced by other schools

with strong and improving test scores in that district. This problem led state leaders to reconceptualize accountability so that it applies to the school, rather than at the district level.

In June of 1989, the Kentucky Supreme Court ruled the public school system in the Commonwealth was unconstitutional. Based on the evidence presented in *Rose v. the Council for Better Education,* (1989), the court concluded that each child in the Commonwealth was **not** being provided with an equal opportunity to have an adequate education. The inequities between rich and poor school districts were too large, depriving children in poorer districts a fair and equal opportunity to receive an adequate education. According to the court, the responsibility for providing an adequate education for **all** children of the Commonwealth rests with the General Assembly. In response to the court order, the state legislature passed the Kentucky Education Reform Act of 1990 (or KERA).

KERA includes a number of legislative mandates, two of which are described here. One mandate is that a primarily performance-based assessment procedure be used. Instead of using only multiple-choice questions as did KEST and CTBS-IV, KERA required the KDE to assess what "students could do with what they know." As a result of this mandate, KDE designed and has been developing the performance assessment component of KERA. This assessment system is named the Kentucky Instructional Results Information System (KIRIS).

KERA also mandated that the assessment system (KIRIS) must be usable for granting rewards to schools that have an increased proportion of successful students and for delivering sanctions to schools that have a decreased proportion of successful students. As documented in Guskey (1994, p.81),

> The legislation requires the State Board to establish . . . a threshold level for school improvement . . . to determine the amount of success needed for a school to receive a reward. The threshold definition shall establish the percentage of increase required in a school's percentage of successful students, as compared to a school's present proportion of successful students, with consideration given to the fact that a school closest to having one hundred percent (100%) successful students will have a lower percentage increase required.

KERA further requires that school success shall be determined by measuring a school's improvement over a two year period. As discussed in Guskey (1994, p. 82), a school that does not reach its prescribed threshold level

> . . . but maintains the previous proportion of successful students shall be required to develop a school improvement plan and shall be eligible to receive funds from the school improvement funds pursuant to KRS 158.805. A school in which the proportion of successful students declines by less than five percent (5%) shall be required to develop a school improvement plan, shall be eligible to receive funds from the school improvement fund, and shall have one or more Kentucky distinguished educators assigned to the school to carry out the duties as described in KRS 158.782. A school in which the proportion of successful students declines by five percent (5%) or more shall be declared by the State Board for Elementary

and Secondary Education to be a 'school in crisis,' and the State Board is to implement defined sanctions.

The rewards and sanctions make the KERA reform a "high stakes" program and KIRIS a high stakes assessment. We note here that the most severe sanction, the school in crisis sanction, has not yet been implemented.

Description of the KIRIS Assessment

In order to understand some of the issues discussed in this evaluation report, it is necessary to have a rudimentary understanding of the elements of the KIRIS assessment program. The purpose of this section is to provide the reader with this rudimentary information.

It should be noted that the basic design of the KIRIS Assessment was established in 1991 in the initial request for proposals (RFP) to develop the assessment system. The Kentucky legislature commissioned a national panel of experts in assessment to develop the RFP. KDE and the winning contractor, Advanced Systems for Measurement and Evaluation (ASME), have had to adapt and implement the legislatively defined design for the KIRIS Assessment.

The accountability index. The basis for describing a school's accomplishment is the KIRIS accountability index. The accountability index for a school is the average performance of the school's students over six separate measures: five cognitive achievement measures and one noncognitive achievement measure. Each cognitive achievement measure reflects a school's performance in one curriculum area. A school's performance on each of the six measures is also reported.

Cognitive achievement at the school level. The measure of a school's accomplishment in each cognitive achievement area is the average achievement score of its students. For each of the five curriculum areas, a student's score is obtained as follows. As a result of several types of assessments in an area (which are described later), each student is classified into one of four quality levels: novice, apprentice, proficient, and distinguished. Next each student is assigned points on the KIRIS score scale as shown in Table 1.

Table 1: Relationship Between Level of Student Performance and Accountability Score Scale Points

Student Quality Level	Corresponding Accountability Score Scale Points
Novice	0
Apprentice	40
Proficient	100
Distinguished	140

A student's score on the cognitive dimensions has a possible range from 0 to 140. If a student is absent from the assessment, the student is assigned a KIRIS score scale of 0.

After assigning points for a curriculum area to each student, all of the students' scores are averaged. The average is calculated separately for grades 4, 8, and 11 (previously 12). The process is repeated for each of the five curriculum areas. These averages are a school's cognitive measures.

A school's noncognitive achievement. In addition to the five cognitive measures, each school receives a noncognitive measure. This measure is a composite of a school's attendance, retention, dropout, and transition to adult life assessments (dropout applies only to middle schools and high schools and transition to adult life applies only to high schools.) The highest possible score a school may attain on the noncognitive assessment is 100.

A school's accountability index. The single number by which a school is judged is the accountability index. This index is the school's average performance over all the cognitive and noncognitive measures. This sum is then divided by six.

Although the theoretical range of the index that combines cognitive and noncognitive assessments is 0 to 133.3, the extremes of 0 and 133.3 cannot be attained except in extraordinary circumstances. For example, in order for a school's accountability index to equal 0, all students in a school must score at the novice level and the school must have a score of 0 on the noncognitive measure. Similarly, at the other extreme all students must score at the distinguished level and the school's noncognitive measure must equal 100. (Interested readers may find a more complete description of the computation of the accountability index for the first biennium (1992-1994) in Guskey [1994].)

The desired minimum score for a school. Based on what we were able to learn from reviewing documents, field work, and feedback from state officials, we believe there is confusion among stakeholders concerning what minimum score is desired or required by each school. Our understanding of the position of KDE staff is that they believe the KERA legislation does not define a "successful student" as a student at the proficient level in one or all subject areas. Rather, they have operated on the assumption that the State Board for Elementary and Secondary Education has the authority to define school and student success. This may explain why the board established four performance levels rather than the one success level that is mentioned in the legislation. Although many Kentucky educators have talked about all schools reaching the 100 level within 20 years, KDE staff do not see this as a firm requirement of existing legislation or regulation.

However, according to the 1991-92 Technical Report, it is our perception, and one that seems to be widely shared by educators in the field, that achieving at least the proficient level for all students is a school's goal. This qualitative goal can be translated to a quantitative accountability index value. Since the proficient level translates to 100 on the KIRIS score scale, an average student score of 100 in each cognitive area is the desired minimum accountability score. We perceive that each Kentucky school is to reach the desired minimum of 100 in 20 or fewer years.

How a school receives rewards or gets sanctioned. Based on the 1991-92 KIRIS assessment, schools received a baseline score on the accountability index. The baseline score was subtracted from 100 (the desired minimum accountability score 20 years after the start of the program). The difference between the desired accountability score (100) and the baseline score was a gap that each school had to close. The gap between the desired accountability score and the baseline score was divided by 10. The division by 10 represents the length of the program--10 bienniums, or 20 years. The gap divided by 10 represented the average gain on the accountability index needed by a school to avoid sanctions.

An example may clarify the previous paragraph. Assume a school received an accountability score of 40 on the baseline assessment. This baseline score (40) would be subtracted from 100. The difference between the desired accountability score (100) and the baseline score (40) is a 60 point gap. This gap would be divided by 10, to account for the length of the program. In this case, the gain on the accountability index needed by this school to avoid sanctions is 6 points. The school's 1992-93 and 1993-94 KIRIS accountability index results would be averaged to determine the school's accountability index value at the end of the first biennium.

If the school in this example had a biennium accountability average of 46, the school would neither be sanctioned or rewarded. If the school's accountability index average was 47 or higher, the school would receive financial rewards. If a school's accountability index average was less than 40, the school would face sanctions under the KERA legislation. If the school's average accountability index was between 46 and 47, the school would be classified as successful. Such a result would subject the school to neither rewards or sanctions. The KERA legislation did not clearly define what happens to a school such as this one if its average accountability index value was between its baseline and its threshold (in this example, between 40 and 46). The KDE has determined that if a school does not achieve its threshold (46 in this case), but increases its accountability score, that school needs to develop a school improvement plan.

How the Assessment Tasks Within a Cognitive Measure are Weighted

Since the initial year (1991-1992) cognitive measures were obtained in the curriculum areas of mathematics, reading, science, social studies, and writing. Each area is assessed by a variety of formats that are weighted differently. In 1993-1994, the formats for the first biennium were weighted as shown in Table 2.

Table 2: Cognitive Weights on the KIRIS Assessment 1992-1994

Assessment Format	Component		
	Social Studies Science Math	Reading	Writing[1]
1. Open-Ended Common Questions (Five Questions)	40%	50%	NA
2. Open-Ended Matrix-Sampled (Each student is randomly assigned to answer 2 of a pool of 24 questions.)	40%	50%	NA
3. Performance Events	20%	NA	NA
4. Multiple Choice Questions[2]	0%	0%	NA
5. Portfolio[3]	0%	0%	100%
6. On-Demand Writing Prompts[4]	NA	NA	0%
TOTAL	100%	100%	100%

How the Components Within the Noncognitive Measures are Weighted

The weights of the components comprising the noncognitive measures are shown in Table 3.

[1] Writing is assessed only through a portfolio.

[2] Multiple-choice tests were administered, but the results were not counted in the accountability measure.

[3] Only mathematics was assessed by a portfolio, and the results were not counted toward the accountability measure.

[4] On-demand writing prompts were administrated, but the results were not counted in the accountability measure.

Table 3: Noncognitive Weights on the KIRIS Assessment, 1992-1994

Component	4th grade	8th grade	High School
Attendance	80%	40%	20%
Retention	20%	40%	5%
Dropout	NA	20%	37.5%
Transition to Adult Life	NA	NA	37.5%
TOTAL	100%	100%	100%

It is important to note that starting with the 1994-1995 KIRIS assessment, the noncognitive measure will be lagged by one year. That is, the computation for the 1994-1995 school year will be based on a school's 1993-1994 data for the four components. This was done because there was insufficient time to collect and disseminate the data for the 1994-1995 school year.

Recent Changes in the Assessment

Because the KIRIS assessment is an innovative and developing program, it is reasonable to expect that changes and fine-tuning will be done each year the program is in place. The following changes are those of which we are aware for the 1994-1995 school year.

1. Assessment using a mathematics portfolio at the fourth grade will be discontinued and in its place, a fifth grade mathematics portfolio will be used.

2. Whereas the mathematics portfolio has not been counted in the accountability index in the past, it will count in the 1994-1995 assessment. The exact weight of the mathematics portfolio will be set at the State Board for Elementary and Secondary Education meeting in February 1995.

3. The accountability grade in high school will be grade 11 instead of grade 12. However, the portfolios in writing and mathematics will still be due in grade 12.

4. Although the 1994-1995 accountability grades will be 4, 8, and 11, grade 12 students will also be assessed. This is for purposes of equating scores from previous years to the current year.

5. Although curriculum areas beyond the five mentioned previously were assessed in 1993-1994, they did not count in the accountability index. These areas are arts, humanities, and practical living/vocational studies. These assessments will count in the second biennium. Currently, arts, humanities, and practical living/vocational studies are assessed in the scrimmage (practice) tests that can be administered in nonaccountability grades. Performance events were also used for these areas in 1993-1994.

Evaluation Questions

Although KIER suggested an extensive list of evaluation questions, addressing the total set was beyond the constraints of available time and resources. Accordingly, we addressed the following subset of questions:

Consistency with Legislative Mandate

1. Is KIRIS consistent with the legislation?

    a. What were the legislative requirements?
    b. To what extent has the legislation been fulfilled?

Clarity of the KIRIS results to users

1. To what extent do various stakeholders understand, at a level appropriate to their role, the assessment and accountability components of KIRIS?

2. To what extent do various stakeholders view the assessment and accountability components of KIRIS as fair, reliable, and valid?

Involvement of Stakeholders in Designing and Development of KIRIS

1. To what extent has there been and is there currently sufficient representation of key groups (e.g., KDE, teachers, administrators) in the development of KIRIS?

Accuracy, Accessibility, and Clarity of Documentation

1. To what extent are the testing materials and instructions that go along with the tests technically sound and "user friendly?"

2. To what extent is the documentation describing the assessment system including the technical reports, accurate and user friendly?

Impact of KIRIS Accountability Policies

1. To what extent is KIRIS perceived by stakeholders as fair to schools with high percentages of minority and low SES students?

2. Will the KIRIS definitions of a successful and an unsuccessful student likely have a positive, an adverse, or no effect on special populations?

3. Is the KIRIS system of rewards and sanctions likely to produce a positive or desirable net effect on the quality of education in all Kentucky schools for all the students?

4. Does the KIRIS effectively promote long-term, positive effects on motivating teachers and students to improve teaching and learning processes?

Technical Adequacy of the KIRIS Assessment

1. The Accountability Index

   a. Does KIRIS produce aggregated indexes of student performance measures sufficiently valid for the intended uses?

   b. To what extent is the KIRIS practice of establishing an initial school baseline of student performance measures a sufficient means of taking into account differences in student backgrounds (e.g., SES and education levels of parents)?

2. Longitudinal vs. Cross-Sectional Accountability Data

   a. Are the KIRIS cohort comparisons preferable to longitudinal comparisons--what are the pros and cons of each?

3. Quality of Technical Information Supporting KIRIS

   a. To what extent does KIRIS meet currently accepted technical standards appropriate for high stakes assessments, statewide assessments, performance assessments, and portfolio assessments?

Evaluation Procedures and Sources of Evidence

During the period of May through December 1994, our team evaluated KIRIS by using the following procedures: (1) collected and reviewed most of the relevant technical documentation of KIRIS; (2) collected and reviewed reports, newspaper articles, monographs, etc., having to do with KIRIS; (3) observed administration of the KIRIS performance events in a school; (4) interviewed key figures involved in the development of KERA and KIRIS; (5) interviewed key figures involved in the implementation and use of KIRIS; (6) observed meetings of Kentucky educators concerned with KIRIS; (7) conducted informal discussions with several groups, including teachers, parents, principals, superintendents, and district assessment coordinators; (8) surveyed selected groups--legislators, parents, district assessment coordinators, and superintendents--to obtain their views on KIRIS; and (9) developed draft reports and subjected them to review by experts on education in Kentucky.

Appendix A contains a list of all main steps involved in conducting this study. In addition, a technical report documenting our sampling framework, response rates, questionnaire forms, and results of our surveys has been provided to KIER.

Limitations of Our Study

This evaluation was conducted under key constraints that undoubtedly have limited what we were able to learn about KIRIS. These limitations should be kept in mind as the reader evaluates this report and decides how best to use it in assessing KIRIS. KIRIS is a hugely complex assessment program that is still under development, and this report was prepared at a much too early time in its short life to warrant its use as a summative assessment of the KIRIS.

We had to conduct this evaluation before KDE had sufficient opportunity to obtain and report some key technical data that we needed to assess the validity of KIRIS. The limited budget and time line under which we worked only allowed us to survey school districts, not schools. Additionally, we were able to visit only a small number of schools while conducting field work in Kentucky.

In spite of the limitations of our study, which must be acknowledged, we obtained a rich set of information and perspectives from a wide range of sources. If carefully considered and weighed, we believe that our conclusions and recommendations should be useful for illuminating and addressing in a constructive way the many issues in KIRIS that are still unresolved.

Answers to Evaluation Questions

Consistency with Legislative Mandate

1. Is KIRIS consistent with the legislation?

   a. What were the legislative requirements?

   The Kentucky Education Reform Act of 1990 fundamentally changed public education in the Commonwealth in a number of different areas. One of the most significant changes was to initiate a school accountability program based on the assessed performance of students. The assessment component has two distinctive features: It is performance-based and it is implemented in a "high stakes" setting. High stakes means that significant rewards are given to schools that meet or exceed specified reward criteria, and sanctions are imposed if the school fails to meet the specified minimum performance criteria.

   Other specifications of this legislation include the following:

   The State Board for Elementary and Secondary Education shall be responsible for creating and implementing a statewide, primarily performance-based assessment program to ensure school accountability for student achievement of the goals set forth in KRS 158.645. The program shall be implemented as early as the 1993-1994 school year but no later than the 1995-1996 school year. The board shall also be responsible for administering an interim testing program to assess student

skills in reading, mathematics, writing, science, and social studies in Grades four (4), eight (8) and twelve (12). The tests shall be designed to provide the state with national comparisons and shall be the same as, or similar to, those used by the National Assessment of Educational Progress. The interim testing program shall begin during the 1991-1992 school year and shall be administered to a sample of students representative of each school and the state as a whole. The test scores shall be used, along with other factors described in KRS 158.6451, to establish a baseline for determining school success during the 1993-1994 school year.

By July 1, 1993, the State Board for Elementary and Secondary Education shall disseminate to local school districts and schools a model curriculum framework that is directly tied to the goals, outcomes, and assessment strategies developed pursuant to this section and KRS 158.645 and 158.6453. The framework shall provide direction to local districts and schools as they develop their curriculum. The framework shall identify teaching and assessment strategies, instructional material resources, ideas on how to incorporate the resources of the community, a directory of model teaching sites, and alternative ways of using school time.

In addition to statewide testing for the purpose of determining school success, the Board shall have the responsibility of assisting local school districts and schools in developing and using continuous assessment strategies needed to assure student progress.

b. To what extent has the legislation been fulfilled?

As can be seen from the preceding excerpt from the legislation, KERA called for extremely broad and sweeping educational reforms. Since many of the mandated reforms require innovation for which off-the-shelf programs did not exist, the KDE was given a most challenging task and a short time to complete it. KDE had to (1) develop the logistics concerning the new assessment (open-ended questions, performance events, portfolios); (2) train teachers to prepare students to submit portfolios; (3) develop a system to train teachers to evaluate portfolios; (4) set standards without a history of data; (5) define successful students; (6) develop guidelines for the participation of special education students; (7) determine the weights for the accountability index; (8) establish thresholds for all schools; (9) develop a rewards system; (10) define the criteria for sanctions; (11) respond to a continuing flow of inquiries from educators and the public and generally keep the Kentucky stakeholders informed; and (12) deal with other requirements presented by the legislature. It should have been expected, therefore, that the Department could not implement all reforms immediately and had to choose an appropriate priority sequence in which to phase in the mandated reforms.

On the major issue of performance-based, high stakes assessment, KDE has pursued the intent of the legislation. The Department was required to produce a fundamentally

different kind of assessment for Kentucky students than the previously used state assessment tests (KEST and CTBS-IV). Additionally, KDE implemented an assessment designed to deliver rewards and sanctions to schools based on defined performance criteria, as explained previously in this report. With the KIRIS assessment, KDE produced a procedure broadly consistent with legislative mandates.

KDE should be complimented for the great strides it has made in acting on the KERA mandates. At the same time, it should be recognized that it is necessary for KDE to continue developing the program in order to attain the vision the legislature set forth. This will require further product development and creating the sophisticated technological basis for demonstrating the validity and effectiveness of the innovations. Examples of what we conclude will need to be accomplished are described below.

1. KDE has succeeded in implementing nonmultiple-choice assessments in the mandated curriculum areas. What remains to be done is to move forward with the performance-based or "doing" aspects of assessment the legislature envisioned. One aspect of performance-based assessment remaining to be more completely implemented is assessing students' abilities to perform nonpaper-and-pencil tasks. Currently, students are required mostly to complete paper-and-pencil tasks. Also, in our judgment KIRIS would be strengthened by including assessments of students' abilities to work cooperatively and productively in groups. (We understand that KDE has concluded that it would have to obtain explicit legislative authorization to assess students in this area.)

2. KDE has succeeded in implementing one of the two components of continuous assessment presented in Appendix B. The legislation mandates the Department to assist schools in "developing and using continuous assessment strategies needed to assure student progress." One way to assist schools in this regard is to provide them with model assessments on which to practice. KDE, through the Advanced Systems in Measurement and Evaluation (ASME) contract, has created practice tests called "scrimmage tests," modeled after the summative evaluation format of the KIRIS assessments. Scrimmage tests have been prepared for the nonaccountability grades (i.e., grades K-3, 5-7, 9-11). What remains to be done is to assist schools to develop the instructionally-embedded continuous assessments envisioned by the developers of the curriculum frameworks. At the moment, these more formative evaluation assessments may be developed by some teachers without KDE's assistance. The quality of the scrimmage tests and the teacher-made continuous assessments should be evaluated. The results of the evaluations should be used to help classroom teachers to improve both the tests and their use of the test results.

3. KDE has progressed rapidly in the assessment arena during the present transitional period. The legislature stipulated 1992 to 1996 as the transitional period. Among KDE's accomplishments is effectively addressing the administrative and educational

issues involved in moving from a multiple-choice testing system toward the ultimate performance-based system.

During the interim testing program, KERA required the State Board for Elementary and Secondary Education to develop and employ tests to provide the state with national comparisons. The legislation requires that the tests be NAEP-like, i.e., the same as or similar to those of the National Assessment of Educational Progress (NAEP). It should be noted, that "NAEP-like" assessments is not a well-defined term. Currently, NAEP uses a mixture of multiple-choice, constructed-response, and performance assessment tasks in a low-stakes assessment setting. Further, the NAEP historically has used matrix sampling for the explicit purpose of preventing comparisons of schools and districts, which conflicts with Kentucky's need to assess progress at the school level. KDE needs to obtain clarification of the legislature's intent concerning national comparisons and then work out the indicated technological basis with respect to comparing the progress of Kentucky's students with that of students in other states. KDE has provided national comparisons for two subject-matter areas in the 1992-93 Technical Report. We understand that KDE plans to issue additional comparisons of the KIRIS results with the NAEP results when future NAEP results become available.

4. While Kentucky has made significant progress in designing and implementing performance-based assessment, there is still much room for improvement. KDE made important progress in directly assessing students' abilities to perform relevant nonpaper-and-pencil tasks on the mathematics portfolio, although this assessment did not count for accountability purposes in the first biennium. A writing portfolio was created by KDE and is counted in the assessment. Beyond this, KDE has done little of this type of assessment. KDE reported to us that it and the State Board for Elementary and Secondary Education have recognized this limitation. It reported that at its September 1994 meeting, the State Board for Elementary and Secondary Education approved KDE's recommendation to embark on an 18-month research and development effort to expand the number and kind of item types used in KIRIS.

5. KDE has also been successful in designing and implementing a school-level accountability index designed in the spirit of the legislation. What remains to be addressed, however, is the reconciliation of KDE's accountability index with the technically different index envisioned by the legislation. There, school sanctions would not be applied unless there was a significant drop in the percentage of successful students. "The percentage of successful students" is technically different from "numerical value of the accountability index" (as it is currently calculated). The following examples illustrate the consequences of following KDE's accountability index as contrasted with the legislation's percent of successful students rule.

Suppose a school's accountability index drops from 41 to 35. This 6 point drop would place the school "in crisis." (We assume, for the sake of this example, that the schools in crisis component of KERA was implemented). However, suppose the percentage of proficient students in the school dropped from 18 percent to 15 percent. A court might well have to decide whether a 6 point drop in the accountability index could place a school in "crisis" if the percentage of proficient students dropped by less than the legislatively mandated 5 percent.

A second example is a school that improves its percentage of proficient students by a sufficient amount to be eligible for rewards calculated by the KERA, but does not receive any rewards because the increase in the accountability index is not sufficient. A court might then have to determine whether the accountability index is sufficiently consistent with the legislation that mandates accountability based on the percentage of proficient students.

Additionally, a school could increase its score on the accountability index by a sufficient amount to be eligible to receive rewards, yet not receive rewards because KDE requires schools to reduce the percentage of novice students by 10 percent over a two-year period to receive rewards. This additional stipulation is not included in the KERA legislation.

KDE has reported to us that all regulations are reviewed by a legislative committee to determine if they are in conflict with either the spirit or the letter of the law. KDE reports that the legislative committee decided there was no conflict between KDE's administrative regulations regarding the calculation of the accountability index and the law. However, it is not clear whether this review process is sufficient to overcome possible challenges to the legality of KDE's accountability scheme. We recommend, therefore, that KDE seek legal counsel to advise it on the review process.

Understanding and Confidence of Stakeholders in the KIRIS Assessment

1. To what extent do various stakeholders understand, at a level appropriate to their role, the assessment and accountability components of KIRIS?

KERA is a complex piece of legislation that accomplished more than simply mandating a performance assessment system. As envisioned by its supporters, KERA would increase parental involvement in the education of their children. Teachers, principals, and parents would work together in the Site-Based Management Councils (SBMC) to improve instruction. It is important for stakeholders to understand the philosophy on which KERA was based. Based on our review of the KIER study done by Wilkerson and Associates, Inc. (1994), we conclude that there are (at least) eight propositions that constitute the philosophy of KERA:

A. All children can learn at a relatively high level.
B. The state should set high standards of achievement for all children.
C. More learning resources should be focused on students who have not succeeded in meeting the state's learning standards.
D. Decisions affecting instruction can best be made at the local school level.
E. In the primary schools, students should not be labeled as belonging to a specific grade level.
F. It is not enough to require that students show their knowledge of facts; they must also demonstrate that they can apply what they know in real life situations.
G. Both rewards and sanctions are necessary to hold schools accountable for improving student performance.
H. Higher performance levels by all children are important for economic growth of Kentucky.

KDE has made significant progress in helping stakeholders to understand the assessment and accountability components of KIRIS. Based on all that we could learn from focus groups, surveys, interviews, and relevant reports, we conclude overall that most stakeholders have some understanding of the rewards and sanctions component of the KIRIS assessment. However, KDE should continue to educate all stakeholders on the specifics of the rewards and sanctions. These specifics are probably known only to a limited number of stakeholders (KDE personnel, superintendents, and district assessment coordinators [DACs] in local school districts, some teachers and principals, state legislators sitting on accountability committees, and some ASME employees.)

We believe that the lack of understanding among a broad group of stakeholders will inhibit acceptance of the KIRIS assessment system. Therefore, we recommend that KDE continue to emphasize effective stakeholder education efforts to improve stakeholder understanding of the assessment and the associated rewards and sanctions.

District assessment coordinators deal with KIRIS on a day-to-day basis in local school districts. Although superintendents are at least one step removed from the day-to-day operational issues of the KIRIS assessment, they represent an important other group of stakeholders. We developed surveys to ascertain the opinions of both these groups on both KERA and KIRIS and attempted to obtain responses from every DAC and superintendent in the Commonwealth. For the DACs, 113 out of 184 responded, for a response rate of 61.4 percent. For the superintendents, 70 out of 176 responded, giving a response rate of 39.7 percent. Based on our review of the quantity and quality of responses to the open-ended and fixed response questions, we inferred that the responding DACs are especially well informed about KIRIS and that the responding superintendents think they are well informed on most of the questions we asked. Clearly, the DACs are better informed than the superintendents on the technical details of KIRIS, which is to be expected. The DACs and superintendents both thought that few or no parents understand the accountability index. Two-thirds (67 percent) of the superintendents and 62 percent of the DACs responded that "all or most" of the parents do not understand the

accountability index. There was only one superintendent and two DACs who responded that "all or most" parents in their district understood the accountability index.

We asked parents attending a PTA Leadership Conference in Louisville how well they understood the accountability index. We obtained survey responses from 63 of the approximately 600 members of PTA leadership councils from throughout the Commonwealth who attended this conference. From the DAC and superintendent surveys we expected a small percentage of parents to report they understood the accountability index. This is **not** what we found. Forty-five percent of the parents reported that they understood the accountability index well; a third (32 percent) reported they understood it poorly. However, parents attending the PTA leadership conference cannot be viewed as a representative sample of parents of students in Kentucky public schools. About 80 percent of the parents responding to this survey reported having at least some college, and more than 50 percent reported having a college degree. These parents represent a degree of educational attainment quite atypical of Kentucky parents generally.

Parental knowledge of the accountability index may be related to a school district's SES. We asked the DACs whether they thought parents understood the accountability index. Clearly, we cannot use their responses to generalize from school district characteristics to the knowledge of individual parents. However, as shown in Table 4 below, we found that a higher percentage of DACs from districts with low percentages of students on free and reduced lunch reported that parents in their districts have some knowledge of the accountability index[5].

Table 4: Relationship Between Districtwide SES and DACs' Perceptions of Parental Understanding of the Accountability Index

Category	Low free and reduced lunch	Middle free and reduced lunch	High free and reduced lunch	Number of Dacs
Some parents understand	44.4%	39.4%	35.3%	42
All or most parents do not understand	55.6%	60.6%	64.7%	67
Total %	100.0%	100.0%	100.0%	
Number of DACs	9	56	44	109

[5] A school district with a percentage of students receiving free and reduced lunch less than 20 percent was operationally defined as low for the purpose of this table. A school district with a percentage of students receiving free and reduced lunch between 20 percent and 50 percent was operationally defined as middle for the purpose of this table. A school district with a percentage of students receiving free and reduced lunch over 50 percent was defined as high for the purpose of this table.

We suggest that KDE increase its efforts to educate parents about the meaning of the accountability index. Using the existing documentation, we estimate that it would take someone with a special concentration in statistics and measurement about four hours to understand the assessment component of KIRIS. A more typical individual would likely need more time to understand it. Individuals not comfortable with numbers and spreadsheets might easily get frustrated trying to understand the various scoring and weighing schemes. Therefore, special educational efforts will be needed to explain details of the program well enough so that all stakeholders understand it.

2. To what extent do various stakeholders view the assessment and accountability components of KIRIS as fair, reliable, valid, and legitimate?

It is typical that stakeholders question the validity of new methods to demonstrate accountability. Initial skepticism is an appropriate response because it allows stakeholders to pose questions, seek rationales, and buy time until the new system provides data addressing legitimate concerns. This questioning demonstrates that stakeholders are taking an innovation seriously. Therefore, even though KDE has made great efforts to have stakeholders involved in the KERA implemen ation process, it is not surprising to find that stakeholders have serious concerns.

A recent study conducted by Wilkerson and Associates, Inc. (1994) for KIER found that principals, coordinators/supervisors, teachers, school council parents, public school parents, and the general public **all** ranked student performance on the KIRIS as the measure least likely to provide a reliable indicator of student learning. These diverse constituencies had most confidence in the percentage of students who finished high school. A study of the Kentucky state legislature found that 44 percent of the responding legislators said the most common complaint mentioned by the public was that the KIRIS was an inaccurate measure of students' abilities (Horizon Research International, 1994). A KIER survey of Kentucky superintendents revealed that the validity and reliability of KIRIS was an area of concern and a very high priority for 91 percent of the superintendents (KIER, 1994). Additionally, superintendents thought that establishing the validity and reliability of KIRIS was the most effective action to address their concern about the KERA (KIER, 1994).

As noted above, we obtained survey responses from 63 of the approximately 600 members of PTA leadership councils from throughout the Commonwealth who attended a PTA leadership conference in Louisville. These responses revealed that a majority (56 percent) of those parents have **little** or **no** confidence in the KIRIS assessment and (62 percent) thought that KIRIS does not assess basic skills. (As stated previously, PTA conference attendees are not necessarily representative of the typical Kentucky parents.) According to the Wilkerson and Associates, Inc. (1994) survey report for KIER (p.4):

When asked to rate six different measures that would reliably indicate that students were learning and schools were improving, the six target groups (principals, coordinators/supervisors, teachers, school council parents, public school parents, and the general public) had highly diverse opinions and all groups indicated they had less confidence in the state's testing program, KIRIS, as a reliable indicator of performance than in any of the other measures tested (the percentage of students who finished high school; student scores on standardized tests in mathematics, science, social studies, and reading compared with students nationwide; scores on college entrance exams such as ACT and SAT; employer reports on how well high school graduates of local schools are prepared for the world of work; and teacher reports to parents on how much their students have learned).

The KIRIS assessment has limited use for assessing the educational progress of individual children. Parents (typically) want to know how **their** child is doing in school in relation to the child's own capabilities and sometimes in relation to peer groups in other communities or in other states. The 1992-1993 Technical Report (released in August 1994) states "the current reliabilities are not sufficiently high to make student level decisions without additional information. The KDE is examining this issue to determine how to proceed." Based on the evidence available to us at this time, we concur. KDE may wish to consider the individual vs. school level scores as a policy issue and suggest solutions for ways schools might provide parents with reliable individual-level student scores.

When considering the value of KIRIS to the Kentucky education system, it is important to consider costs. The Commonwealth of Kentucky pays about $6 million a year to ASME alone for developing and scoring the KIRIS assessments and assisting the Department to train Kentucky educators to score portfolios. Many districts pay $7.25 per student to ASME to cover grading costs for the continuous assessment. Professional development costs to teachers constitute another large cost of the program, estimated by the KDE in a letter to us at $2,000,000 per year. The annual cost for the rewards component of the assessment is an additional sum of money, estimated by KDE to be about $18,000,000. We would concur with correspondence received from KDE indicating that it would be appropriate to invest about one percent of the Commonwealth's education budget on "a system that simultaneously is driving student learning to new heights and evaluating the level of learning." However, such an investment could not be justified if stakeholders' current worries that KIRIS is failing in areas of legitimacy, validity, reliability, and fairness are warranted.

In sum, **all** the cited evidence suggests stakeholders have questions concerning the legitimacy, validity, reliability, and fairness of the KIRIS assessment. We have no evidence to suggest that parents think the assessment component of KIRIS is a fair, reliable, and valid system.

This is a very significant finding in view of the large annual investment of time and money required to develop, maintain, and employ KIRIS. It seems clear that KDE needs to develop and disseminate evidence that will convince stakeholders that KIRIS has sufficient merit to be worth what it is costing.

Involvement of Stakeholders in the Design and Development of KIRIS

1. To what extent has there been and is there currently sufficient representation of key groups (e.g., KDE, teachers, administrators) in the development of KIRIS?

KDE made efforts to have representatives of key groups sit on advisory committees (21 content advisory committees, crossing 7 content/discipline areas and 3 school levels). For example, as described in the 1991-1992 Technical Report, advisory committees were established for reading, mathematics, science, and social studies. These committees met 5 times from September 1991 through February 1992. Representatives from KDE, teachers, administrators, curriculum coordinators, and representatives from the Kentucky Education Association were included on the committees. In the second and third years, additional committees were formed to deal with diversity and other areas.

Nevertheless, many teachers are unaware of the input Kentucky teachers have had through these committees. During one of the focus group inquiries we conducted, one teacher commented that there was no teacher input on the questions used in the KIRIS assessment. Another teacher in the group disagreed saying that she in fact sat on one of the committees. The point of this anecdote is to highlight that there is usually little communication among teachers about their involvement in the development of KIRIS. Despite the committee system and the input of Kentucky educators into the review process of the KIRIS assessment, many teachers perceived that questions on the assessment were constructed by outsiders with little or no knowledge of Kentucky.

Since KERA was a legislative action, it is not surprising that teachers think they are not involved. It is difficult, as the KDE has come to learn, to have sufficient grass-roots involvement when implementing a top-down program. We suggest, however, that KDE continue and strengthen its efforts to involve all teachers in the KERA reforms and the KIRIS assessment specifically. The KIRIS portfolio assessment is one way in which KDE has begun to do this. Continued improvements in its use of portfolio scoring as a teacher development effort should be encouraged. KDE might also consider involving all teachers in submitting tasks (items) for the KIRIS accountability assessment. Even though teachers' items would need to be revised and edited and only a few teachers' items would actually appear on an assessment, the task development exercise would help teachers focus on performance-based assessments and allow them some direct input. It might also be a mechanism to facilitate the use of continuous, formatively oriented performance assessment in the classroom.

Accuracy and Accessibility of Documentation

1. To what extent are the testing materials and instructions that go along with the tests technically accurate and user friendly?

A review of the KIRIS assessment and the scrimmage tests revealed that the instructions, assessment materials, layout, and presentation of material were straightforward. Additionally, the DACs and teachers did **not** report that students had a problem understanding the directions on the KIRIS assessment or the scrimmage tests. For these reasons, we conclude that the testing materials and associated documentation are user friendly.

We did not study the accuracy of each question or each scoring rubric. However, on the whole, we found the KIRIS and the scrimmage assessments to be technically accurate. We know of only one reported instance of a technically inaccurate multiple-choice question asked on the 1991-1992 KIRIS exam (see Appendix C). Since the multiple-choice items did not count on the KIRIS assessment, this item was of no consequence in determining a school's accountability index.

2. To what extent is the documentation describing the assessment system, including technical reports, accurate and user friendly?

The rapid way in which KDE chose to create and implement the KIRIS assessment resulted in some good practices and some problems. On the positive side, some apparently interesting and curricularly faithful assessment tasks were produced and administered to students statewide. At the same time, this created masses of data, ripe for technical analyses and documentation, that were not analyzed appropriately. It is typical that an education agency that moves rapidly to an annual operational assessment program underestimates the amount of human energy and fiscal resources it must invest in the production phase of the program. This means that the technical phase, which provides policy makers with the necessary information to suggest the validity and reliability of the assessment results, is usually neglected or given low priority. KDE appears to be no exception in this regard.

A great amount of data on the KIRIS assessments is not reported in the technical reports. Additionally, technical information on KIRIS is spread among many documents and analyses that are not reported in the technical reports. Some of these documents address technical information well, but others do not. It is difficult for anyone reviewing the program to grasp the technical basis for the program by reading these scattered documents. A technical report should completely and accurately present the technical basis to enable an outside reviewer to evaluate the reliability and validity of the assessment program's results.

The contents of an assessment program's technical reports should follow the guidelines suggested by the Standards for Educational and Psychological Testing (APA, 1985).

These Standards, developed by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education, are recognized throughout the assessment profession as appropriate for all assessment programs. Further, the Standards have been used in some legislation and in the courts to support or refute testing programs. Recent technical writings on performance assessment (e.g., Linn, Baker & Dunbar, 1991) extend or expand on the Standards rather than replace them. Similarly, Linn (1994) has recently argued that performance-based assessment programs that include performance standards need to provide validity evidence that focuses on the intended uses of the assessments by the state and the positive and negative consequences occurring to students, teachers, and the educational system if those intended uses were implemented.

The KIRIS technical reports do not seem to have been organized from the perspective of laying out clearly for an outside reviewer what data support and what data refute the use of the KIRIS results in the way they are intended to be used. Neither the perspective of the Standards nor the perspective of the performance-based extensions articulated by Linn, Baker and Dunbar (1991) and their associates are found in the technical reports.

Tables 5, 6, and 7 provide suggested areas around which the KIRIS technical reports might be organized. The tables' contents were specified from the perspective of the Standards and the Linn, Baker and Dunbar (1991) extensions described previously. They are intended to represent the topics that the professional assessment community would consider important to include in a technical report so that outside reviewers can evaluate the technical merits of the program. Emphasis is on organizing data to support the way KDE intends the KIRIS results to be interpreted (i.e., constructs and their intended meanings) and to support the use of the KIRIS assessments in light of intended and unintended consequences (i.e., consequential validity). The organization of the tables seems entirely consistent with performance-based assessment programs intending to serve accountability purposes. Nevertheless, KDE may wish to have other assessment specialists outside of Kentucky review any outline for a proposed revision of its KIRIS technical reports.

We note that some of the suggestions in Tables 5, 6, and 7 may require KDE to conduct special studies outside the context of the operational assessment program. Conducting such studies is typical professional practice in the assessment development areas. In some important areas, it may not be possible to obtain the necessary scientific support of KIRIS using only the data produced by the operational assessment program.

We note here that the excellent start made by KDE on the KIRIS assessment program could be undermined by failing to provide relevant technical information. Critics of programs quickly point to insufficient or inadequate documentation of validity and reliability. This provides grist for encouraging those who doubt the program. Thus, a good technical basis for a program is necessary not only for the technical community but also to support and sustain policy.

We understand that KDE has not had sufficient time and resources to fully document and report on the soundness of KIRIS. However, KIRIS is already being used to reward schools and to provide extra aid for schools not meeting the defined performance threshold requirement. Based on communications with KDE, we were informed that the Biennium I Technical Manual is due out in early February 1995 and that much of the information cited in Tables 5-7 will be included in that report. KDE's plan to release a full technical report soon is most timely and important.

Impact of KIRIS Accountability Policies

1. To what extent is KIRIS perceived by stakeholders as fair to schools with high percentages of minority and low SES students?

The evidence on this question is from interviews with teachers and administrators in Owensboro, Bowling Green, and Jefferson County school districts plus the responses of superintendents and the DACs to one survey question. Based on this limited evidence, there appears to be some concern about the fairness of KIRIS to schools with high percentages of disadvantaged students.

The DACs and superintendents report that low SES students--presumably disadvantaged students--have more difficulty with written communication. Teachers in the focus groups pointed out that there are questions on the KIRIS assessment that may not be appropriate for low SES students. For example, one question on the KIRIS assessment asked students what they would like to do (or what they did) on vacation. Two teachers independently mentioned that low SES students may not have gone on vacation and may not be able to understand the concept of a vacation with sufficient accuracy to write about it.

2. Will the KIRIS definition of a successful/unsuccessful student likely have a positive, adverse, or no effect on special populations?

The limited evidence we gathered with respect to special populations showed that the educational system was paying considerably more attention to these groups than before the KIRIS system was set up. Special education students always take the KIRIS assessment unless they are severely handicapped. A severely handicapped student is assessed using an alternative portfolio. From the focus groups we concluded that special education teachers have occasionally been pleasantly surprised at how well their students did on the assessment. Some educators reported in the focus groups that the educational system was paying more attention to special education students because such students were included in the accountability system. Once again, there is **no** evidence that students have been stigmatized by KIRIS. If KDE can document that special education students are receiving increased instructional attention because of KIRIS, this evidence would support the consequential validity of the accountability index.

Table 5: Example of a KIRIS Technical Report Chapter Outline for Reporting Reliability and Validity Data for the School Accountability Index

I. Distribution Statistics

A. Distribution of School Accountability Index for each component and for total composite (including frequency distributions means, standard deviations, etc.)

B. Geographic/regional distribution (including descriptive statistics)

C. Distributions by school size (including descriptive statistics)

II. Stability of Accountability Index Over Time

III. Construct Validity of the Index

A. Survey data showing that teachers, principals, and school boards understand the meaning of the index

B. Survey data showing that leading curriculum experts agree on the instructional usefulness of the index

C. Observational data showing differences in what actually happens in the classroom for those schools with high vs. those with low values of the index

D. Factor analysis of the subcomponents of the index along with other school variables to show the structure of the data

IV. Consequential Validity

A. Survey data to show that the desirable outcomes of using the index and accountability system are being attained and the degree to which they are attained

B. Survey data to show that no serious undesirable consequences of using the index have resulted

C. Evidence that schools with high concentrations of minorities and poor students are not adversely affected by using the index

Table 6: Example of KIRIS a Technical Report Chapter Outline for Reporting Reliability Data For Assessment Components

I. Reliability and Standard Error of Measurement (SEM) (student-level data)[6]

A. Open-ended questions (math, science, reading, social studies)

1. Internal consistency/generalizability, reliability, and SEM
2. Scorer reliability, percent agreement, decision consistency, and SEM
3. Equivalent forms and SEM (e.g., two consecutive years)
4. Test/retest, stability over time, and SEM

B. &

C. Portfolio (writing and mathematics)

1. Internal consistency/generalizability, reliability, and SEM
2. Scorer reliability, percent agreement, decision consistency, and SEM
3. Equivalent forms and SEM (e.g., two consecutive years)
4. Test/retest, stability over time, and SEM

D. Performance events

1. Internal consistency/generalizability, reliability, and SEM
2. Scorer reliability, percent agreement, decision consistency, and SEM
3. Equivalent forms and SEM (e.g., two consecutive years)
4. Test/retest, stability over time, and SEM

E. Multiple-choice items (if included in future assessments)

1. Internal consistency/generalizability and SEM
2. Test/retest, stability over time, and SEM
3. Equivalent forms and SEM

---

[6] Although student level assessment data are not now encouraged for interpretative purposes, there may be increased pressure to do so. Reliability data at the student level may be used to help make policy decisions regarding the desirability of pursuing student level assessment interpretations.

Table 6, Continued

F. Full-test reliability (open-ended and performance)

1. Internal consistency/generalizability and SEM
2. Test/retest, stability over time, and SEM
3. Equivalent forms and SEM (e.g., two consecutive years)

G. Summary of reliability and SEM in relation to decisions and interpretations made of the results

II. Reliability and Standard Error of Measurement (SEM) (school-level data)

A. Open-ended questions (math, reading, science, social studies)

1. Generalizability of school means
2. Sources of error variance in school means

a. Items/tasks
b. Students
c. Occasions
d. Others

3. Standard error of measurement and standard error of means for school means
4. Stability of school means organized by school size and by initial level of school performance
5. Decision consistency of school means
6. Reliability of school mean change scores (interpreted in terms of school growth or progress)

B. &
C. Portfolio (Writing and Mathematics)

1. Generalizability of school means
2. Sources of error variance in school means

a. Items/tasks
b. Students

Table 6, Continued

c. Occasions
d. Others

3. Standard error of measurement and standard error of means for school means
4. Stability of school means organized by school size and by initial level of school performance
5. Decision consistency of school means
6. Reliability of school mean change scores (interpreted in terms of school growth or progress)
7. Effect of teacher training on reliability of school means

D. Performance events (all subject areas)

1. Generalizability of school means
2. Sources of error variance in school means

a. Items/tasks
b. Students
c. Occasions
d. Others

3. Standard error of measurement and standard error of means for school means
4. Stability of school means organized by school size and by initial level of school performance
5. Decision consistency of school means
6. Reliability of school mean change scores (interpreted in terms of school growth or progress)
7. Effect of teacher training on reliability of school means

E. Noncognitive Assessment

1. Report the consistency and accuracy with which scores report/collect information regarding the components.
2. Report the stability of the noncognitive measure (and its components) over 1, 2, and 3 year periods.

Table 7: Example of a KIRIS Technical Report Chapter Outline for Reporting the Validity of the KIRIS Tests

A. Representativeness of Academic Expectations

1. Blueprint showing match of items to academic expectations (AE)
2. Table showing summary of formal ratings of advisory committees of tasks for curricular match, relevance, importance
3. Other evidence

B. Construct Interpretation

1. Correlations among the scores within and between subject-matter areas for each assessment component (multiple choice [MC], open-ended, performance, portfolio) (student level and school level data)
2. Evidence of improvement following specific teaching practices (student level data)
3. Correlations with CTBS for all components (MC, open-end, performance, portfolio) (student level data)

C. Differential Results (within and between components and for each subject)

1. Male vs. female
2. Grade-level changes
3. Geographical differences within Kentucky
4. Ethnic differences, poor vs. nonpoor differences
5. Relation to teachers' years of experience (the above are student level data analyses)

D. Consequential Validity

1. Positive impact on teaching practices
2. Negative impact on teaching practices
3. Impact on student learning and attitudes

3. Is the KIRIS system of rewards and sanctions likely to produce a positive/desirable net effect on the quality of education in all Kentucky schools for all the students?

We understand that KDE has no choice about whether or not to provide rewards to schools that exceed their defined threshold on the accountability index. These rewards were mandated by KERA.

It is too early in the reform movement to predict with any confidence whether rewards and sanctions will be beneficial or detrimental to educational quality in the Commonwealth of Kentucky. Rewards from the first biennium are slated to be distributed to teachers by March 1995. Stakeholder groups are split on the potential impact of rewards and sanctions. Teachers and principals are opposed to the concept of sanctions (Wilkerson & Associates, 1994).

Our surveys of DACs and superintendents provide some evidence on this question from two other stakeholder groups. Only 14 or 20.3 percent of the 70 responding superintendents and 30 or 26.5 percent of the 113 responding DACs thought that eliminating the threat of sanctions based on the accountability index would improve instruction. On the other hand, 33 or 47.8 percent of the 70 responding superintendents and 48 or 46.2 percent of the 113 responding DACs thought eliminating the threat of sanctions based on the accountability index would not improve instruction. The rest of both groups did not know or were undecided.

From the DAC and superintendent surveys, we also found that a majority of respondents thought that eliminating rewards from the accountability index would not improve instruction. Only 10 or 14.5 percent of the superintendents thought eliminating rewards from the accountability index would improve instruction. For the DACs, only 14 or 12.5 percent thought eliminating rewards from the accountability index would improve instruction. On the other side, 42 or 60.9 percent of the 70 responding superintendents and 65 or 57.5 percent of the 113 responding DACs thought eliminating rewards from the accountability index would not improve instruction.

Clearly, the above-referenced surveys give some evidence to support the use of rewards and sanctions to obtain a positive influence on education in Kentucky. This is corroborated by the survey for KIER by Wilkerson and Associates, Inc. (1994). They found that the majority of responding public school parents and the general public thought that rewards and sanctions are necessary to hold schools accountable for improving student performance, while the majority of responding principals and teachers thought that rewards and sanctions are not necessary to hold schools accountable for improving student performance. Also, another superintendent survey conducted by KIER in conjunction with the Kentucky Association of School Superintendents (KIER, 1994) included a question on rewards and sanctions that was worded differently and showed that superintendents thought rewards and sanctions are **not** necessary to hold schools accountable for improving student performance.

It seems fair to conclude that the issue of rewards and sanctions based on KIRIS results is controversial. Below we give our analysis of what the positive and negative influences of the rewards and sanctions *might* be.

A. Possible Positive Impacts of the KIRIS Assessment System of Rewards and Sanctions

1. Increased concentration on students' writing. As envisioned by KERA supporters, the KIRIS system of rewards and sanctions is supposed to motivate principals, teachers, and site-based management councils to alter the instructional curriculum presented to students. It has been argued in Kentucky that "assessment drives curriculum." The KIRIS assessment is heavily oriented toward writing. The biggest weight on the KIRIS assessment is assigned to the open-ended questions. The second biggest weight is assigned to the writing portfolio. In the five cognitive areas covered in the first biennium, writing in one form or another accounted for 88 percent of the weight on the KIRIS assessment. Counting performance events as writing (students have to write their responses to situations) increases the weight of writing to 100 percent of the KIRIS assessment. Even in subjects like mathematics, students must write answers to questions. Students who do not write well will not do well on the KIRIS assessment, irrespective of their knowledge of subject matters like mathematics. Such a heavy weight on writing on the assessment may have made it easier for educators to stress writing when teaching. Students do report more writing under the reforms (Coe, Leopold, Simon, & Williams, 1994).

2. Improvement in students' writing quality. Oral reports from KDE also indicate that scores on the writing portfolios have improved. Teachers, DACs, and superintendents report almost unanimously that writing has improved; and the writing improvement was over and above what would have been expected of most school children of the same age. We *believe* (but we cannot be sure) that the reported increase and improvement in students' writing is due to the heavy weight on writing on the KIRIS assessment and, ultimately, the prospect of rewards and sanctions based on KIRIS assessment results. It is a limitation of our study that we did not gather and study student portfolios and other evidence to assess whether the quality of students' writing has actually improved.

3. Involvement of students in cooperative problem-solving. Performance events add a group component to the KIRIS assessment. However, performance events accounted for 12 percent of the weight of the KIRIS assessment in the first biennium. Even with a relatively light weight, the performance events (group activities involving problem solving or experimentation) meaningfully engage students. Consistent with the 1994 revised legislative requirements, they yield experience but not assessments of the ability of individual students to work productively and collaboratively in groups and perform important learning targets.

Students reported increased group work since the passage of the KERA (Coe, Leopold, Simon, & Williams, 1994).

4. Instructional contribution of portfolios. Another benefit of the KIRIS assessment has been the development of portfolios. Portfolios of students' mathematics and written work appear to have great instructional potential. Students have to choose their best work to put in the portfolio, and this gives students a chance to reflect upon their intellectual growth over a school year. A more passive assessment system, like a test consisting of 100 percent multiple-choice questions, does not meaningfully engage the student in the same way. Choosing pieces for inclusion in the portfolio engages the student more, thereby increasing the student's involvement in deciding the material on which the student is to be assessed. For teachers, evaluating the best work of **their** students gives them critical feedback that they may use in deciding what instructional materials and assistance would best serve the entire class as well as individual students.

B. Possible Negative Impacts of the KIRIS Assessment System of Rewards and Sanctions

1. A crisis of trust resulting from scoring and rescoring of portfolios. Portfolios have had at least one negative impact due to an incident involving the scoring and auditing of portfolios. KDE requires teachers to score their own students' portfolios because this provides insights to help guide classroom instruction. Portfolio scores vary considerably depending on which teacher did the scoring, making portfolio scores less reliable than the open-ended writing questions. This makes the auditing of portfolio scores necessary to ensure accuracy and fairness. The first such audit, done by ASME in 1993, showed a serious level of unreliability. Out of the 105 schools selected for inclusion in the audit of the writing portfolio, 99 schools were found to have significantly higher scores by the original teacher than either the Kentucky summer rescorers or the New Hampshire audit scores.

   This finding had a strong negative impact on teacher morale. Problems in scoring the portfolios have to some extent eroded stakeholder confidence in the reliability and validity of KIRIS. In our focus group meetings, teachers and administrators reported very negative reactions to the audit. The teachers reported that the state seemed to suspect them of either cheating or incompetence. They said that morale among the affected teachers was very low. This was supported by our survey of DACs. Forty-eight out of the 54 responding DACs indicated that teacher reaction to the audit was negative. Only 5 (9.3 percent) said that teacher reactions were neither positive nor negative. In response to many complaints from the field, KDE gave schools 6 options concerning which scores should be used for the writing portfolios (1992-1993 Technical Report, Ch. 7, p. 20.)

2. Overconcentration of assessment based on writing ability. The KIRIS assessment currently is heavily oriented toward evaluations based on writing. Students who have content knowledge of the discipline but lack adequate writing skills are precluded from doing well on the assessment. We understand that KDE is conducting research to develop performance assessments that can be evaluated by using modes of expressing knowledge beyond writing. Based on our conversation with Mr. Richard Hill (President of ASME), some examples of alternative kinds of assessments could include oral communication, involving the giving of a speech; creating a presentation that simultaneously involves written and visual information, typically called multimedia and usually done on a computer; and performing and fine arts. We note too that multiple-choice items offer students who lack adequate writing ability an alternative way to express the knowledge they have. Multiple-choice tests require good reading skills, however.

4. Does KIRIS effectively promote long-term, positive effects on motivating teachers and students to improve teaching and learning processes?

The Commonwealth just completed its first biennium in implementing KERA. While it is appropriate to search for effects, it is too early to expect definitive information on results. It is also important to consider what the long-term effects are likely to be given the extent of changes in instruction and assessment that are occurring.

KIRIS undoubtedly is motivating some teachers and students to improve teaching and learning, especially in the area of writing. Also, the guiding legislation intends that teachers at all grade levels should use instructionally embedded performance assessments to help guide student learning. The state accountability system is being designed to provide a model of high quality, performance-based testing. As conceptualized by proponents of KIRIS, there would be two continuous assessment strands: a formal assessment, called scrimmage tests, and a set of instructionally embedded assessments crafted by the teachers. The formal scrimmage tests would be developed by the contractor and made available to local school districts. The state's assessment is supposed to provide inspiration for teachers to develop their own performance assessments. We could find no evidence to suggest that any significant proportion of teachers is using the KIRIS or scrimmage tests to develop their own instructionally embedded assessments. To realize the aim of involving teachers at all grade levels in instructionally embedded performance assessments, the state will have to determine how to increase the interest and involvement of teachers in the assessment development process, especially those in the nonaccountability grades. Earlier we suggested asking each teacher to contribute performance-based tasks to the assessment task pool. This activity may have the effect of teachers crafting such tasks for use in their own classrooms.

From the focus groups conducted in Louisville, Bowling Green, Owensboro, and Bell County, we concluded that fear has motivated many teachers and principals to consider altering their practices in teaching and administration. Fear primarily stems from the

most severe sanction of KERA, the school in crisis designation. Educators are afraid of losing their jobs. As the Commonwealth of Kentucky acquires more experience with the KIRIS assessment, educators will probably learn that the likelihood of a school going into crisis is low. Thus, the consequences of sanctions, especially the most severe sanctions, may not be as dire as educators originally thought. With a reduced fear factor, teachers and principals might not pay as much attention to the KIRIS system of sanctions.

Vermont implemented a low stakes statewide portfolio assessment program during the 1991-92 school year. Teachers involved in the Vermont portfolio assessment reported portfolios as a worthwhile burden (Koretz, Stecher, Klein, & McCaffrey, 1994). However, these same teachers reported that portfolios caused considerable stress. Koretz, Stecher, Klein, and McCaffrey (1994) report

> The pressures experienced by educators went beyond time demands. For example, more than half reported difficulty finding appropriate tasks. Educators also reported feeling stress because of their uncertainty about appropriate uses of portfolio scores; the rapid implementation of the program; and inadequate, tardy, and inconsistent information from the state.

In focus groups held in Kentucky, we found that teachers thought that portfolios were beneficial to instruction. However, Kentucky teachers reported the same concerns as their Vermont counterparts. Increased stress was reported by nearly every teacher attending the focus groups. Additionally, a study of one Western Kentucky school district found that teacher stress is extremely high, approaching debilitating levels for many (Hughes & Craig, 1994). From our focus groups we concluded that teacher stress was especially strong for teachers involved with the accountability grades.

Portfolios increase teacher workloads. This increase in teacher workload occurred in both low stakes assessment programs (Vermont) and a high stakes assessment program (Kentucky). It is *possible* that Kentucky teachers experience greater job stress than Vermont teachers due to the high stakes Kentucky assessment. However, we have no data on this question. We conclude that some of the job stress reported by Kentucky teachers is due to the assessment system and not to the accountability provisions of KERA.

Technical Adequacy of the KIRIS Assessment

1. The Accountability Index

   A. Does KIRIS produce aggregated indexes of student performance measures that are sufficiently valid for the intended uses?

      We looked at several aspects of the accountability index to answer this question. As is typical of most systems that try to reduce complex outcomes to a single

dimension, the KIRIS accountability index is difficult to interpret meaningfully. It is affected by a number of factors that complicate its meaning. Some of these factors are reviewed below.

Size of School May Affect the Index's Consistency

Small schools usually will have larger fluctuations in the accountability score, no matter what the teachers do. If there is a group that comes through the school with 20-30 students in the accountability grade, and 5-6 are dramatically different from last year's class, the overall accountability score will change. Schools have no control over this situation. This problem is reduced, but not eliminated, in middle size schools. Larger schools have fewer problems with the consistency of the index.

KERA has recognized and addressed the possibility of problems with fluctuations in the accountability index due to factors other than school effectiveness. The index is based on two years of assessment data, which should minimize this problem for all but the smallest schools. Also, if a school's decision-making body believes its accountability index decreased due to factors outside its control, it can enter an appeal asking the state to review and possibly revise the accountability index. According to KERA (Guskey, 1994, p. 84), "The state board may adjust a performance judgment on appeal when evidence of highly unusual circumstances warrants the conclusion that the performance judgment is based on fraud or a mistake in computations, is arbitrary, is lacking any reasonable basis, or when there are significant new circumstances occurring during the biennial assessment period which are beyond the control of the school."

It is important for KDE to study the issue of within and between school fluctuations in the accountability index. The KDE should monitor index fluctuations in relation to school size, grade level, school SES levels, and geographical region. Data on the sizes of these fluctuations would be helpful in interpreting reliability and validity studies and may also inform policy decisions related to the index. These variability studies should be reported in the Technical Report.

Changes in Teacher Retention Rates May Affect the Index

There are hosts of potential factors that can affect a school's instructional program. As a consequence, the accountability index can go up or down. One determinant of doing well on the open-ended tasks may be a teacher's familiarity with open-ended tasks and knowing the type of instruction and practice that students need. When several new teachers come into a school and do not give the type of instruction and practice that would increase student performance on such tasks, scores on the accountability index are likely to be affected. The new teachers might be very good, but might not be familiar with the type of instruction and practice that maximize student performance on KIRIS. This conclusion represents the collective judgment

of the evaluation team. We do not have any Kentucky-based evidence that bears directly on this potential effect.

Data on teacher turnover nationally have shown that urban schools that predominately serve low SES students are likely to have higher rates of teacher turnover than schools that serve a more affluent clientele (North Central Regional Education Laboratory, 1995). If the same phenomenon occurs in Kentucky, schools that predominately serve a low SES clientele may be unfairly sanctioned. We recommend that KDE conduct a study to investigate the relationship among teacher retention rates, SES of students, and accountability scores after the completion of the first biennium.

Measurement Error Affects the Accountability Index

Kentucky must take every precaution to insure that KIRIS evaluations of schools are based upon reliable information. It would be grossly unfair to administer rewards and sanctions based on inconsistent estimates of school effectiveness. Thus, KDE must provide solid evidence that substantiates the reliability of the accountability index.

The Department reported impressive reliabilities for the school accountability index in the KIRIS 1992-93 Technical Report. Across grade levels and sizes of school, the reported reliabilities are all .90 or above. However, these reliabilities were calculated using both students and assessment items as fixed, not random factors. Consequently, the generalizability analyses cited in the 1992-93 Technical Report (Ch. 9, p. 5) may yield a higher reliability estimate than would have been found with alternative assumptions. Had students "been considered a random source of error, the generalizability indices reported here would probably be significantly lower, since the variation of students within a school is considerable." Also, the rescaled item scores may yield somewhat smaller variances than the actual item scores may be expected to yield, since the regression procedure used to rescale the item scores may reduce the individual differences among the originally assigned scores on the items. As stated in the 1992-92 Technical Report, "this variance [of students within a school] would have a substantial impact on small schools, where differences among students from year to year may be large" (1992-93 Technical Report Ch 9, p. 5). Consequently, it is difficult for us to understand and evaluate KDE's efforts to assure that the accountability index is sufficiently reliable to inform policy, administer rewards, guide school improvement efforts, and administer sanctions.

We recommend that KDE calculate the generalizability coefficients and variance components assuming both students and assessment items as random sources of error. These can be reported along with the now reported results. While this dual reporting would not resolve the debate about which model is the more appropriate, it would show readers the consequences to the reliability estimates of using one model or the

other. We continue to believe that students and items should be considered as random sources of error in the generalizability model employed, since scores from one set of students and one set of assessment items are obtained in one year to set a threshold for evaluating the performance of a different set of students on a different set of assessment items in a subsequent year. Further, policymakers are concerned with the effectiveness of future cohorts of students on yet-to-be-created assessment tasks. They are interested in the current cohort not only in their own right, but also because this cohort signals the future trends of the educational system. Thus, policymakers should be informed of the reliability of the index as it is affected by student and item sampling.

We also recommend that KDE engage one or more psychometricians who are nationally known specialists in generalizability theory to (a) review KDE's current generalizability data collection design and analyses and (b) assist KDE in formulating a new design if it is necessary to do so. The specialists should also address appropriate ways to report standard errors and variance components estimates.

We do recognize that the most important question about reliability concerns reliability at the school level. There is not sufficient evidence in the available technical reports for us to assess whether the accountability index is or is not a reliable measure of a school's effectiveness in meeting KERA standards. The KIRIS assessment tasks have great value for the instructional process. However, their value for holding schools accountable is not as great because there is no clear agreement that the school level assessments are sufficiently reliable for this purpose.

While the main concern is not about the reliability of student level data nor about the reliability of the components of the accountability index, it nevertheless seems appropriate to comment on these. Overall, there is a considerable amount of measurement error associated with individual student level scores. For this reason, the 1992-1993 Technical Report (Ch 9, p. 10) states that " . . . current reliabilities are not sufficiently high to make student-level decisions without additional information. The Department of Education is examining this issue to determine how to proceed. One of two responses is anticipated for the 1994-1995 assessment; either to increase the number of open-ended items or to include multiple-choice data in the calculation of students' total scores." The ACT study reaches the same conclusion: "ACT strongly advises that no judgments regarding individual student decisions can or should be made at this time on the basis of the present Kentucky performance test results" (ACT, 1994).

Measurement error is high for two of the three components of the KIRIS assessment: portfolios and performance events. The rater reliability of portfolio scores for writing was shown to be low, when ASME rescored portfolios for 105 schools and found that "99 had differences outside the acceptable scoring range (1992-93 Technical Report). As also reported in the 1992-93 Technical Report, "The

reliability estimates of the performance events were highly unstable. They varied from 0.00 to 1.00."

Measurement error is much lower for the open-ended questions. Student level reliability on open-ended questions varied between .69 and .87, depending upon grade and subject matter (1992-93 Technical Report, Ch. 9, p. 10). The reliabilities are higher in the 12th grade than in the 4th grade.

The reliability results found in the 1992-1993 Technical Report are based (mainly) on one component of the assessment, open-ended questions. The considerable amount of measurement error in the portfolio assessment may decrease the legitimacy of KIRIS with key stakeholders, especially teachers.

Validity of the Assessment System for Improving Instruction

An accountability system should provide teachers with guidance on how to change classroom activities in the context of teaching that is unique to a particular school. There is much in the system designed to assist school staffs to examine and change school and classroom instructional practices. Some of these include academic expectations, curriculum framework, units of study, training sessions, scrimmage tests, released assessment items, content area school indices, portfolio scoring analysis reports, instructional implications documents, distinguished educators, regional consultants, and Kentucky Education Television (KET) broadcasts. Also, the assessment tasks may be good examples of learning targets for students.

Validity of the Accountability Index for Improving Instruction

In contrast to the other components of the assessment system, the accountability index itself does not provide feedback that is meaningful and constructive to teachers. A good example of appropriate feedback is that given to teachers with the writing portfolio after teachers became upset when their original ratings were lowered. The feedback was that teachers were weighing the surface features of the writing too heavily and the deeper, more substantive issues of the writing too lightly. Feedback should include not only what is needed to increase the scores, but also what needs to be taught. This type of feedback with respect to mathematics, social studies, and science would be very helpful. Just a number (the accountability index) does not tell the teacher very much about how to improve.

Reporting Progress on the Goals and Academic Expectations is Missing

In addition, the accountability index does not report (separately) progress on each of the Kentucky education reform goals. Neither does the accountability index report progress on easily understood groups or clusters of specific academic expectations. The components of the index are linked to traditional, fragmented, and

compartmentalized subjects (English, math, etc.). Thus, the accountability index and its components do not seem to be in the spirit of those provisions and descriptions laid down early in the reform program.

Index subscores should be developed to report on the progress of students meeting each of the four reform goals. Each question on the KIRIS assessment should be identified as covering a goal or set of goals. The index should give a report on progress toward each of the four goals. Clearly, the goals overlap and development of an index to report on the goals would not be an easy task, but neither would it be impossible. If the goals are to be meaningful, even if there is an overlap, it is important to report on them.

The Index Lacks Policy-Making Utility

A single index of educational quality is not sufficient to guide policy decisions about schools.

Policymakers need information they can use to make decisions about parts and levels of the educational system that need to be improved. For example, they need to identify and assess accomplishments in given subject matter areas and in relation to the reform goals and for different levels and categories of students. Policymakers need to know whether or not individual subjects are being neglected and whether all areas within a subject are improving. They need to know what the educational gains are for students in different parts of the state, students at different grade levels, and students with varying needs. We understand that KDE is conducting some of these analyses presently. We hope that KDE will disseminate the results from these analyses in one publication, such as the upcoming Biennium I Technical Report.

By itself, the accountability index provides only limited information for state-level educational policy-making. However, the data underlying the index are a rich source of policy-relevant information. KDE seems to be using these data usefully to help policymakers examine educational policy issues. For example, they reported that the standard deviations increased at all grade levels between the 1991-92 and 1992-93 assessment years. However, much more could be done if sufficient resources were directed toward using the KIRIS data to inform policy-making.

So far, the Department has not reported data on some of the relevant policy issues. These include achievement of the 4 reform goals, student performance on the 57 academic expectations, comparisons of the performance of advantaged and disadvantaged students, comparisons of the achievements of ethnic groups, and comparisons of the performance of males and females.

We recommend that KDE use Table 7 above to decide with policymakers what supplementary analyses should be provided to accompany the information on the

accountability index. We also recommend that KDE consider developing subindexes covering such variables as subareas within a subject, male and female achievement, and achievement of the main ethnic groups in the schools. Such subindexes should be useful to schools as well as policymakers for diagnostic and planning purposes. KDE may also wish to consider allowing educational policy researchers in the academic community to have access to its database. These researchers could have the time to conduct studies that the KDE staff may not have the time to do.

The Index Does Not Take Into Account Factors Beyond a School's Control

The index is vulnerable to influences of extraneous variables, and the accountability system includes two important safeguards against such influences.

Policymakers and school personnel need unambiguous information about the effectiveness of schools. This information should indicate what the school is accomplishing in improving student learning apart from factors not under its control. Before making high stakes decisions about a school, state policymakers and school councils need to be certain that accountability scores are not influenced by conditions **not** under the school's control. It would be inappropriate to sanction a school if its low score on the accountability index were due to insufficient resources or changes in the student population.

The accountability system seems to acknowledge this in that it provides two important safeguards against judging schools on factors not under their control: averaging two years of the KIRIS accountability index and the KERA provision that a school's staff can appeal if it believes its effectiveness was unfairly assessed and judged.

We agree with KERA that schools where students are underachieving should be treated appropriately, whatever the reason for the poor achievement. However, we believe that it would be an invalid use of the accountability index to sanction school personnel if the poor achievement of their students is due to factors beyond the control of the teachers, administrators, and school council. Such factors might include student mobility and socioeconomic change in the community. The personnel in such schools should be assisted to improve student learning but should not be stigmatized and otherwise penalized. The provisions for averaging accountability index results over two years and the appeal mechanism somewhat mitigate this concern but do not alleviate it completely. Thus, we recommend that KDE study and report on whether changes in student mobility and neighborhood socioeconomic conditions are associated with schools receiving rewards and sanctions.

B. To what extent is the KIRIS practice of establishing an initial school baseline of student performance measures a sufficient means of taking into account differences in student backgrounds (e.g., SES and education levels of parents)?

The KIRIS practice of establishing an initial baseline using the single year of data from the 1991-92 school year does not sufficiently account for differences in student backgrounds. If the baseline established for a school is unduly influenced by factors not under the school's control, then the subsequent accountability assessment two years later may provide an invalid estimate of what the school has accomplished. Thus, the appeal provision of the KERA is important for correcting erroneous judgments not only about the two year's growth in student achievement, but also the school's baseline accountability score.

One method for explaining the inadequacy of the current approach is to look at the link between socioeconomic status and change in accountability scores. Over a 20 year period, schools can serve a fundamentally different clientele. Some areas will see their economies grow. Schools in these districts will see the socioeconomic status (SES) of their students improve. Meeting (or exceeding) numbers required by an accountability index will be relatively easier for these districts than for schools experiencing SES decline. Two areas visited in Kentucky showed signs of such growth: one is the "golden triangle" (areas in northern Campbell, Kenton, and Boone Counties), and the other is eastern Kentucky. However, some schools in Louisville would show a decline in student SES if, as we predict, the African-American and white middle class families move away from the central city. Schools in these areas would serve a fundamentally different clientele in future years than that on which the prior baseline data are based. For these schools, the teachers might be doing an excellent job teaching students. However, the measure of the job done by the teachers is limited to the KIRIS accountability index. In these schools the accountability scores would go down if the SES base of the students decreases. Because SES is related to the performance on the KIRIS assessment, teachers in those districts might realize that their chances for monetary rewards are reduced, no matter how hard they work, and their likelihood of receiving sanctions is increased. (Some teachers have already come to that conclusion.) It is not necessary to know exactly how student populations will change in different schools to make this claim against the validity of the accountability index. Our argument on the point requires only the assumption that changes in student populations do occur in some schools and that such changes are bound to affect the accountability index. Thus, the initial baseline is not sufficient to take into account differences in student backgrounds.

The second way to look at the inadequacy of the current approach is to consider the students in a school that primarily serves low SES children. Schools that serve students in the inner city often serve a clientele of transient students. Even if instruction was working to improve assessment scores, it may not be reflected in the accountability index, especially if new students coming into the school were

academically weaker than those leaving the school. Thus, the initial baseline is not sufficient to take into account differences in student backgrounds.

A counterargument to taking students' background into account is that schools need to adjust to economic changes just like businesses. That is, the "real world" of business and industry cannot hide behind conditions that are out of their control, but are held accountable to turn a profit or get out of business. From this perspective, businesses either adapt to economic change in a community, move to a different location, or they go out of business. However, a business is different from a school in several important respects. Businesses have considerable control over their raw materials. Public schools cannot "choose" their students, but must do the best they can with whatever "material" (students) show up at the door. Also, a business can cut out or eliminate unprofitable parts of the business. A school cannot "cut out math," for example, if it is having difficulty with students learning the subject. Further, businesses are often in a position to make rapid changes and see immediate results. Schools cannot respond that rapidly. For example, if fourth graders are poor readers in 1994, improving the 1995 fourth grade reading program will not help immediately. Schools **must** improve the 1st, 2nd, 3rd, and 4th grade reading programs since reading is cumulative and develops slowly. Businesses can also relocate to a place with a more favorable economic climate. If a city is in economic decline, a business can leave the city and relocate to the suburbs. A **public** school is fixed in a community and cannot pick up and leave one community for another where resources and raw materials are better. Schools have to adapt to changing economic conditions without the exit option available to business and industry.

Schools in areas of declining SES may be able to adapt to a changing economic base in the community and might be effective in teaching a new base of students. However, under the present accountability system, current students will be compared with prior students. If those two groups of students differ in their SES and the older group has, on average, a higher SES than the newer group, the school faces an increased risk of falling into the sanctions category.

The impact of the relationship between change in SES and change in accountability is more likely to evolve over a longer period than a biennium. Over a two-year period, **most** communities are sufficiently stable so that ignoring changing SES is an acceptable action. In the unusual case of rapid SES change, we would recommend that the State Board for Elementary and Secondary Education use its power listed in Section 5, subsection 8 of KERA. The pertinent section of the paragraph reads:

> The State Board may adjust a performance judgment on appeal when evidence of highly unusual circumstances warrants the conclusion that the performance judgment is based on fraud or a mistake in computations, is arbitrary, is lacking any reasonable basis, or when there are significant new circumstances occurring during the biennial assessment period which are beyond the control of the school.

2. Longitudinal vs. Cross-Sectional Accountability Data

A. Are the KIRIS cohort comparisons preferable to longitudinal comparisons--what are the pros and cons of each?

This question must be considered in relation to the questions to be answered by the comparisons of student achievement from year to year. We understand that the State Board for Elementary and Secondary Education mainly wants to know whether student achievement at three grade levels is being brought to the proficient level of achievement. If this is all it wants to know, and if it is the only audience for the analyses, then cohort comparisons may be adequate, although they are affected by changes in student population.

We think that educators in the schools are also an important audience for the year-to-year comparisons of student achievement. Based on our surveys of the DACs and superintendents, we believe that educators want to know more than whether educational scores are improving at three grade levels. The DACs and superintendents who responded to our surveys overwhelmingly prefer a longitudinal approach over a cohort approach for the KIRIS assessment. DACs supported a longitudinal approach: 87 percent preferred it, and 5 percent opposed it (with 8 percent responding "don't know"). The superintendents supported a longitudinal approach: 77 percent preferred it, and 16 percent opposed it (with 6 percent responding "don't know").

Based on our interviews and focus groups, we believe that educators in a school prefer longitudinal analyses because they want to know if their students are making satisfactory educational progress from year to year. For example, if a high percentage of students were found to be at the novice and apprentice levels during the fourth grade year, the fifth grade teachers and their colleagues in the school would understandably want to know whether the subsequent year's efforts were successful in helping the underachieving students reach the proficient level of achievement. The cross-sectional approach, which now assesses students only at selected accountability grade levels, provides the teachers in the school and the interested policymakers no help in answering questions about year-to-year gains of a group of students. We believe that longitudinal comparisons would be more useful than cohort comparisons in helping educators determine if a school's instructional program is making a difference in student learning. However, this would only be so if the students were assessed at least every other year, not only at one grade level in each school as now is generally the case.

We realize that the scrimmage tests might assist in meeting the need for longitudinal data. However, the technical adequacy of these tests has yet to be demonstrated. Assuming that policymakers and school personnel are both important audiences and

that there is thus a need to follow students over time, then the longitudinal approach is to be preferred.

Longitudinal analysis would give a better picture of what is happening to a group of students as it goes through the educational system and would provide more information on what levels and aspects of curriculum and instruction need to be improved. Longitudinal comparisons are not administratively easy, include smaller samples of students, and are probably more expensive than cohort comparisons. Since we believe the longitudinal comparisons offer more guidance for improving instruction and since Kentucky educators seem to prefer this approach, we think KDE should seriously examine the possibilities of replacing the cohort approach with the longitudinal approach. The following discussion of this issue may be useful.

Longitudinal comparisons provide a school with more information about where the problems lie. Such comparisons allow one to determine if students left a grade in poor shape and whether they improve later. Conversely, a longitudinal model would be able to assess if students left a grade as proficient in the assessed subjects, but deteriorated later. If students were proficient when they left the 4th grade but did not progress by 8th grade, longitudinal data assessing their achievement levels at the 5th, 6th, and 7th grades would help 8th grade teachers determine what went wrong and at what grade level and in what subject matter areas. As seen in this example, the longitudinal model has one less source of error than the cohort model: the students do not change from "pretest" to "posttest." In the cohort model, some of the students included in the analysis have been in the school less than one year, others have been in the school one full year, and still others two years or more. This situation does not yield as good an indication of what happened to the students between the 4th and 8th grades as would be the case if a longitudinal approach was used where each student included in the analysis had been in the school and assessed during the 6th and 7th grades.

A serious issue for longitudinal analysis across grades is the development of an appropriate growth scale. Since the content and skills assessed at different grade levels are not identical, somewhat different abilities may be assessed at different levels. This makes measuring growth on a common multigrade scale problematic. KDE may need to convene both technical and subject matter panels to advise it on how to design such scales.

In this comparison of analysis approaches, the cohort approach includes a larger sample of students, while the longitudinal approach reveals the learning trend for those students who were in the school for the three assessment years.

For schools with high rates of transient students, it might be necessary to do both cohort and longitudinal analysis. But longitudinal analysis would give a better

picture of what is happening to a group of students as it goes through the educational system.

Longitudinal analyses would not be able to include all students. Correspondence with Neal Kingston of KDE indicates that there are schools for which longitudinal analyses would include fewer than one-fifth of the students. Longitudinal analyses requires at least two data points for each student (Meyer, 1994). Cohort analysis requires only one data point for a given student.

Longitudinal analysis could probably be done with only minor changes in the current system. The 15 digit code now used by ASME (or any other outside vendor running the statistical analysis) could be revised to include a unique number that could be translated by a KDE employee into an individual's student name. In that manner, a test contractor would not have access to a student's unique number, but KDE could link scrimmage tests taken in the 7th grade with the 8th grade KIRIS assessment. This procedure would give KDE and school personnel a much better view of any problems in the educational system.

3. Quality of Technical Information Supporting the KIRIS

A. To what extent does KIRIS meet currently accepted technical standards appropriate for high stakes assessments, statewide assessments, performance assessments, and portfolio assessments?

There is not one set of "currently acceptable technical standards" for performance assessments. However, the technical standards that are most widely accepted by both the assessment profession and the legal system are the AREA, APA, and NCME (1985) Standards for Educational and Psychological Testing. These Standards apply in a general way to all assessment programs, but they are not worded in such a way that their applicability to high stakes performance assessment systems is clear to the casual reader. To help make the Standards applicability clearer, Linn, Baker, and Dunbar (1991) provided a set of criteria for evaluating performance assessments. Linn (1994) made it clear that these extend or explicate the Standards rather than replace them and that the Standards are generally applicable to large scale assessment programs. Later in this section, we use the Linn, Baker, and Dunbar criteria to evaluate KIRIS.

Four formats were used to obtain a school's 1993-1994 KIRIS accountability score: open-ended common questions, open-ended matrix-sampled questions, performance events, and the writing portfolio. However, the open-ended and matrix-sampled questions are the same format of tasks for students. Thus, there are three different formats of tasks on the KIRIS assessment.

Open-Ended Questions

We find that the open-ended questions generally meet currently accepted technical item-writing standards for such questions. In other words, they are well written, appear to assess higher-order thinking skills, and are free from item-writing flaws. They are also scored with adequate reliability. However, they are not all performance tasks.

Writing Portfolios

The writing portfolios also seem well organized and thought out. However, the reliability of the scoring of the writing portfolios is low. The audit of writing portfolios conducted by ASME in 1993 found that school scores for 99 out of 105 schools had differences outside the acceptable scoring range. According to the 1992-93 Technical Report, these schools were not randomly selected to be audited, but were chosen because their writing portfolio scores were substantially different from the school's on-demand writing scores, other content area scores, and/or previous years' portfolio scores. The discrepancies found in the audit may be larger for this sample than for the state as a whole. Nevertheless, the writing portfolio was shown to be unreliable at both the individual and school levels for almost all of the sample of schools included in the audit.

Portfolio reliabilities are lower than open-ended questions for two reasons:

1. Portfolios are less standardized than the open-ended tasks. That is, students differ greatly as to what they include in the portfolio.

2. The portfolio evaluation task for a teacher is more complex than a similar evaluation task for open-ended questions. (Please note that teachers evaluate the portfolios, while ASME evaluates the open-ended questions.)

We concur that portfolios are useful to teachers for instructional purposes. Their utility for purposes of student evaluation and school evaluation is a secondary consideration.

Given the weight of the writing portfolio in the accountability index (16.7 percent), we recommend that the Commonwealth continue to place great importance on the training of teachers to score the writing portfolio. Portfolios appear to have great instructional potential and therefore should continue to be used for their instructional consequences.

Portfolio scores of Kentucky students will contain considerable measurement error over the next few years. ASME, the contractor conducting the portfolio assessment, *may* be able to achieve high scorer reliability from **their** highly trained portfolio

evaluators. However, these evaluators use their skills in portfolio evaluation continuously. Additionally, scorers at ASME can have their work checked and reviewed continuously. In Kentucky, teachers score the portfolios, but that is a small part of a teacher's job responsibilities. Usually, the teachers score the portfolios intensively for only a couple of days each year. After the task is done, the teachers will not score another set of portfolios for another year. It is not realistic to expect that teachers who spend 2-3 percent of their professional time scoring portfolios will be able to score them as reliably as scorers from ASME.

It might be true that teachers who integrate the Kentucky writing standards into their instructional curriculum on a daily basis will come to know the standards so well that they will be able to score portfolios as reliably as ASME's full-time scorers. Pertinent empirical evidence on a statewide basis would be required to confirm such an improvement in the reliability of teacher scoring of the writing portfolio in relation to increased integration of the Kentucky writing standards into a teacher's classroom.

We believe that teachers in Kentucky should continue to score the writing portfolios of their students. We base this recommendation on the instructional value of portfolios and the value of having teachers seriously evaluate the best work of their students. As a consequence of this recommendation, we understand the reliabilities of the writing portfolios probably will be low and the KIRIS accountability index may thus be less reliable than is desirable.

Performance Events

There are two important features of the performance events. First, they provide an opportunity for Kentucky educational authorities to assess the extent to which individual students can perform important learning targets as contrasted with knowing about the learning targets. Second, they provide a potential means of assessing whether individual students have attained cooperative/collaborative skills and learning targets.

Regarding the latter use, we note that KDE's employment of performance events to assess student achievement is consistent with what current Kentucky law allows. That is, the Department cannot assess Kentucky's learning Goal 4 (become responsible members of a family, work group, or community, including demonstrating effectiveness in community service).

Although students carry out performance events in groups, the current procedures do not assess cooperative/collaborative skills and learning targets. After students complete the group activity, they complete a paper-and-pencil "test" assessing their knowledge of the group activities and their conceptual understanding of the principles

underlying the problem the group solved. Oftentimes, an individual student's ability to perform is not directly assessed.

Thus, we conclude that the KIRIS assessment does not take full advantage of the capability of performance assessment methodology. We must point out, however, that because (a) performance events are included in the accountability composite score and (b) there is nationwide educational publicity about the need to use performance assessments, performance assessment activities in Kentucky classrooms have been reported to increase. Therefore, it would be prudent to continue these events.

If the law would be amended to allow KDE to assess students on Goal 4 of KERA, we believe that Kentucky schools could derive greater instructional benefit from performance events. Then it would be possible to use the performance events to assess the individual's ability (a) to work cooperatively/collaboratively in a group and (b) to perform important learning targets as a member of the group. If this recommendation were adopted, assessment designers should review the Kentucky academic expectations to identify those outcomes best assessed by group performance tasks. The assessment program could then focus on increasing performance assessment of Goal 4 outcomes.

The Linn, Baker, and Dunbar Criteria

Linn, Baker, and Dunbar (1991) proposed eight criteria to evaluate performance-based assessments. These criteria represent an application of the widely accepted Standards (1985) to specific technical issues to which state-mandated performance assessment programs should attend. We review the KIRIS assessment in light of these criteria.

Consequences

Linn, Baker and Dunbar (1991) note that "high priority needs to be given to the collection of evidence about the intended and unintended effects of assessments on the ways teachers and students spend their time and think about the goals of education." The **process** of the KIRIS assessment has changed classroom practice (instruction now stresses writing, especially strategies to answer open-ended questions). The change in classroom practices creates the opportunity to address intended and unintended effects of the KIRIS assessment. One benefit proponents of the KIRIS assessment hoped for is improved writing by Kentucky public school students. Our survey of District Assessment Coordinators (DACs) and superintendents revealed that nearly 100 percent of these two stakeholder groups thought that writing had improved. From the focus groups in Louisville, Owensboro, Bowling Green, and Bell County, it was found that nearly 100 percent of teachers reported that writing had improved. More importantly, the writing improvement

noted by teachers was over and above what they would have expected due to maturation of the school children. As acknowledged previously, we did not, however, sample students' writing to determine whether the perceived improvements were actual improvements.

We heard concerns from the focus groups that a considerable amount of instructional time is covered by substitute teachers because of the KIRIS assessment. If the amount of such substitute involvement were substantial, there could be a concern about whether the quality of instruction were suffering.

Based on our survey of the DACs, the typical teacher is out of her or his classroom to receive training in portfolios (mathematics and writing) about 2 to 3 days. In many, but not all districts, teachers are out of the classroom about 2 days in order to score the portfolios. Students spend nearly 1 week, on the average, completing the on-demand writing and performance events parts of the KIRIS assessment. Together, these add to about 10 days or about 6 percent of the student's time in school during a year. We judge that the student's time in completing the on-demand and performance events parts of the assessment is about the same as the time required to complete other comprehensive testing programs, such as CTBS. We also judge that the additional time required to complete portfolios is an appropriate part of the regular instructional schedule.

We acknowledge that some DACs, superintendents, and teachers see the amount of time teachers are away from their students due to the KIRIS assessment to be a potential problem, but so far the involved time seems to be reasonable. The time that teachers spend training and marking the KIRIS portfolios may be useful as inservice education, and examination of student materials is certainly an appropriate part of the instructional process.

Fairness

A study comparing performance on the ACT and KIRIS provides one important piece of evidence on the generalizability of the KIRIS assessment **at the individual student level** (ACT, 1994). The evidence was mixed. Overall, the ACT study found the following:

> It is evident that the performance-based test results of 1992 reveal the typical trends in student performance as noted above; i.e., men scoring higher than women in science and mathematics, Caucasian students scoring higher than minority students, etc. These results, however, subject the Kentucky tests to the same criticisms of 'apparent' bias as are directed at traditional testing methodologies.

We note that the ACT study referenced above was based on the KIRIS assessment results in 1991-92, the first year of the assessment program, the year when the baseline and thresholds were established. It is possible that the relationship between the KIRIS assessment and ACT would be different in the 1993-94 testing cycle. However, such data are not available at this time. We recommend that the Kentucky Department of Education investigate whether there are large gender and ethnic differences on the various components of the KIRIS assessment. If such differences are found, we recommend that a panel be selected to review the tasks and the results to ascertain whether the tasks are fair to these groups. Although the 1991-92 Technical Report says tasks were reviewed before being used, there are no details provided as to the formality and thoroughness of the process. Nevertheless, empirical data should be collected and used to aid in evaluating fairness.

Generalizability Over Groups and Time

Data supplied by Richard Hill (President of ASME) showed that all **individual** level components of the KIRIS test were positively correlated with each other. The correlations varied in strength. Correlations between the writing portfolio and the other components of the KIRIS assessment were much lower than intercorrelations between open-ended, common questions, and the multiple-choice questions. There is substantial **individual level** variability by task on the KIRIS assessment. At the school level, the correlations are stronger.

The 1994 ACT study, using 1991-92 KIRIS data, found that

> As measures specially designed to assess **group** performance, there exists a positive relationship between the Kentucky and ACT test results. The skills being assessed by the Kentucky performance-based tests are consistent with and positively correlated with the skills measured by the ACT assessment. At the **Distinguished** performance level, the performance of students on both the ACT and Kentucky tests is consistent with ensuring student success in college, based upon the validity research conducted by ACT over the last 30 years. However, at the other three performance levels, the current evidence is somewhat suspect because of the nondiscriminating power of the Kentucky test. (ACT, 1994; p.9)

We suggest that the KDE continue to assess the transfer and generalizability of KIRIS assessment, looking not only at comparisons between ACT and the KIRIS assessment results, but also other standardized tests and practical criteria such as job performance of graduates. Longitudinal comparisons between standardized tests and KIRIS assessment results would allow policymakers to review whether scores on tests were increasing simultaneously with increasing scores on the KIRIS assessment.

There is a technical question that concerns the weighing of the open-ended common and matrix-sample questions for the math, social studies, and science cognitive areas. On the 1993-1994 KIRIS assessment, each student took 5 open-ended common

questions, which counted 40 percent on the cognitive dimension, and 2 matrix-sampled open-ended questions (out of a pool of 24 questions), which also counted 40 percent on the cognitive dimension. The equal weighing of the open-ended common and matrix-sample questions is arbitrary, especially when considering the different number of questions that go into each component. In terms of the accountability index, each open-ended, matrix-sampled question answered by a **student** has more than twice the weight of an open-ended common question answered by that same student. Evidence should be presented to determine what impact (if any) the increased weight per question of the open-ended, matrix-sampled questions has in the achievement assessment results and in the accountability index.

As we discussed in a previous section entitled "Measurement Error Affects the Index Construction," the high reliability estimates reported for the accountability index are based on a generalizability analysis that considers students and items to be fixed factors. We again recommend that the Department recalculate these reliabilities under the assumptions that items and students at the school level are random sources of error and that nationally recognized specialists in implementing generalizability be engaged to help KDE to evaluate, and possibly redesign, its procedures for assessing the reliabilities and standard errors of the components and the accountability index.

Cognitive Complexity

In our opinion, **all** components of the KIRIS assessment, including the multiple-choice questions, **require** higher level thinking skills on the part of students. Proponents of the KERA legislation hoped that the new performance-based assessment would require students to critically think through a set of issues before a question could be answered. ASME should be complimented in putting together tests that require the higher level thinking skills envisioned by the KERA legislation.

Content Quality

This issue was not dealt with in our review. It should be noted, however, that the 1991-1992 Technical Report did not report data on the quality of the content. For example, formal evaluations of content quality by teachers, curriculum developers, and content experts could be collected and the results could be reported in the technical reports each year. Appendix C shows one item whose quality was questioned. To our knowledge, this was the only item that was questioned, and scores on it did not count in the accountability index.

Content Coverage

We note here that the technical report on the KIRIS assessment should provide evidence of the content coverage relative to the frameworks. The technical reports

do mention and describe the frameworks, but there are no data reported as to how many KIRIS tasks assess each framework component. We consider reporting this coverage to be very important in helping stakeholders to understand and assess content coverage. We recommend that future technical documentation from the test developer not be accepted without such basic descriptive information. Good assessment development practices require specifications, blueprints, and documentation of content coverage.

In focus groups, teachers reported that the coverage of the common open-ended questions in the KIRIS assessments varies widely from year to year in relation to the academic expectations. Focus groups are not based upon random selections of teachers. Thus, we are unable to generalize to the perceptions of teachers in general. It is possible that the perception noted in the focus groups is true. If so, this is a serious concern since an assessment that covers different things in different years is not very useful to measure growth in student achievement. If the assessments actually do not vary widely from year to year, then it would be important to make sure that teachers and other stakeholders understand that this is so.

As Linn, Baker and Dunbar (1991) noted, there may be a trade-off between breadth of content coverage and some of their other criteria. Additionally, they note it may be one criterion by which traditional (multiple-choice) tests appear to have an advantage over more elaborate performance assessments. As mentioned in the section under consequences, teachers, DACs, superintendents, and parents attending a PTA leadership meeting all thought that breadth of content coverage in instruction was sacrificed to meet the requirements for the KIRIS assessment. However, we did not independently evaluate whether these perceptions by the responding stakeholders were accurate.

Meaningfulness

An important concern is whether the accountability index and its component scores are educationally meaningful. What does a school learn about how to improve itself from knowledge of its KIRIS scores? The assessment reports and reporting schedule are not designed to provide timely, detailed assessment feedback that school staffs can use to diagnose students' learning deficiencies and focus and improve instruction accordingly. Given the high stakes nature of the KIRIS results for schools, the lack of detailed diagnostic information at the school or classroom level could result in a narrowing of the curriculum to those types of tasks and activities likely to appear on the assessment. To some extent this has occurred already. Schools report teachers training their students to respond correctly to short-answer questions. In so doing, they may not appropriately focus on the concepts and learning targets of the framework that underlies these short-answer questions. (We note that an increase in the number of short-answer questions is scheduled for the next assessment.) Schools also report more performance and writing activities.

We recommend, therefore, that KDE continue to engage advisory committees (made up of such stakeholders as school principals and teachers) to help it identify the specific kinds of curriculum-based information that the KIRIS assessment should supply in order for a school and its teachers to have sufficient information to take action to improve students' learning. This panel may, for example, identify specific curriculum subareas that should have scores. It might also identify a reporting mechanism that relates directly to the goals and subgoals in the academic expectations instead of reporting by subject areas alone. The point is, teachers and principals can and should be able to provide valuable advice to KDE on what assessment reporting details would best help them improve education in their schools. In the context of this recommendation, we are advocating that KDE place more weight on making the assessment feedback as timely and educationally meaningful as possible as opposed to providing mainly an accountability judgment.

The KIRIS assessment is unusual in that its stakes differ for teachers and for students. The assessment is "high stakes" for schools. Based on the results of the KIRIS assessment, schools may find themselves in a reward or sanction condition. Teachers and principals may receive bonus money if the school is in a reward condition. On the other hand, under the KERA provision covering schools in crisis (Guskey, 1994; p. 84), Kentucky's distinguished educators "make personnel recommendations every six (6) months on retention, dismissal, or transfer." Thus, teachers *could* lose their tenure and possibly their teaching posts if the school was declared "in crisis" and the Commonwealth of Kentucky fully implemented the most severe sanction in KERA.

However, the assessment has "no stakes" for students. Students may have little (or no) motivation to perform well on the KIRIS assessment. The assessment is not sufficiently reliable to be used at the individual student level for either accountability or diagnostic purposes. Performance on the assessment tasks do not directly impact a student. Due to the small number of tasks on the assessment, the KIRIS assessment results cannot be used to make individual diagnostic placement decisions.

DACs were asked in the survey to estimate the percentage of students in their district who take the KIRIS assessment seriously. DACs reported that 90 percent of the fourth grade students take the assessment seriously (standard deviation of 12 percent). DACs reported that 78 percent of the eighth grade students take the KIRIS assessment seriously (standard deviation of 16 percent) and that 66 percent of twelfth grade students take the KIRIS assessment seriously (standard deviation of 21 percent). The high standard deviations indicate that in some districts DACs reported a much higher (or lower) percentage of students taking the assessment seriously than the average. Under the Kentucky system, school scores on the accountability index are likely to rise (or fall) depending on whether a higher percentage of students take the assessment seriously when compared to a prior year's assessment.

Some districts report an increase in the percentage taking the assessment seriously in the 1992-93 and 1993-94 administration of the KIRIS assessment, compared to the 1991-92 baseline year. It is possible that some of the gains recently reported on the accountability index are based on an increased percentage of students taking the assessment seriously. We recognize that some school districts in Kentucky have recently instituted policies and practices that place KIRIS results on high school transcripts and that all schools are expected to send KIRIS assessment results for individual students to their parents. However, one problem with the KIRIS assessment remains: KIRIS results have little effect on high stakes decisions that affect individual students, e.g., college admission, course grades, promotion, and graduation. Students may thus have little incentive to do well.

Two Additional Criteria Recommended by the Evaluation Team

In addition to the criteria recommended by Linn, Baker, and Dunbar (1991), we have invoked two additional criteria that we believe are important for evaluating the KIRIS assessment.

Multiple assessment formats. We note that it is not the type of assessment activity that is important--for example, it is not performance activities or enhanced multiple-choice questions per se--but whether engaging in those activities leads to clear answers about whether students are learning the targets specified in Kentucky's academic expectations. It is crucial to consider the effect of the KIRIS assessment methodology on the format and content of instruction. As noted above, the use of the short-answer format has caused teachers to teach students how to better answer short-answer questions. The issue is really bigger than this. The choice of assessment strategies should be driven by the known or suspected effects they have on instruction and learning rather than because the strategy is different. There is nothing inherently wrong with the multiple-choice assessment format or any performance-based assessment strategy. It is in the best interests of driving effective instruction to include a combination of formats. For example, the enhanced multiple-choice format can provide some positive features to the assessment tasks (e.g., highly reliable scoring, less dependency on writing, and increased content coverage). We think KDE has needlessly handicapped the KIRIS assessment system by ruling out the use of enhanced multiple-choice test items. The general point is that including a wide range of assessment item formats and broadening the content coverage through inclusion of efficient enhanced multiple-choice items militates against teachers narrowing instruction to one or a few item types and a narrow range of assessed content.

In recommending that KDE increase rather than narrow the modes of assessment employed, we acknowledge that KIRIS is broader than many testing programs that use only one or two modes of assessment. Nevertheless, our point stands. KDE

should employ as broad a range of assessment modes as is feasible in order to enhance construct validity.

Cost, Efficiency, Practicality, Instructional Features (Please note that this set of criteria encompasses, but goes beyond, the Linn, Baker, and Dunbar [1991] criteria of cost and efficiency.)

1. Can the assessment accommodate typical numbers of students? Yes, this appears to have been accomplished in the KIRIS assessment.

2. Is the assessment easy for teachers to use? No, but it must be pointed out that teachers do not use the assessment results directly, and teachers score only one part of the assessment, the writing portfolio. It is possible that principals and site-based management councils use the results of the assessment more than teachers, but no data exist on this.

3. Do teachers agree that the theoretical concepts behind the assessment procedure reflect the key understandings they are teaching? The theoretical concepts behind the KIRIS assessment are stated in the 4 learner goals and 57 academic expectations found in the Kentucky Curriculum Framework. In the focus groups, teachers expressed support for the academic expectations. This was especially true for academic expectations listed under Goal 2 (the content coverage goal). However, teachers expressed concern that the coverage of the academic expectations varied from year to year on the KIRIS assessment. Thus, while the frameworks may underlie the KIRIS assessments, they may not be adequately represented in the year-to-year administration of the assessment.

Concerns About Future Forms Of the KIRIS Assessment

The most recent (1993-1994) KIRIS assessment scheme includes several modes of assessment: extended open-ended tasks, shorter open-ended tasks, multiple-choice tasks, portfolios, and performance events. This diversity of approaches is a strength of the scheme. The multiple modes of assessment approach is supported by educational assessment specialists because it enhances the validity of the results. Validity is enhanced because allowing students increased opportunities to demonstrate their abilities in a variety of ways increases construct representation. Construct representation is increased in two ways: (1) multiple modes allow a student to demonstrate understanding in many different ways, and (2) multiple modes allow for broader and more representative coverage of academic expectations. We think that employment of multiple assessment methods also mitigate against teaching to a narrowly conceived test.

A recent paper (Further Considerations of Issues Related to the Inclusion of Multiple-Choice Items in KIRIS Reported Scores (Kingston, 1994) argues for

narrowing the KIRIS assessment scheme by (a) discontinuing the use of the multiple-choice mode of assessment in calculating the accountability scores, (b) increasing the number of short open-ended tasks, and (c) eliminating or reducing the number of the longer, more extended open-ended tasks. In addition, the paper does not argue for increasing the number of performance tasks that will comprise the accountability score.

In our view, the paper's recommendation that the Commonwealth of Kentucky stop administering the multiple-choice tasks and thus not count students' performance on them toward the accountability score is problematic. (The State Board for Elementary and Secondary Education decided in September to stop administering the multiple-choice items.) We suggest that the Board reconsider its decision. Below are brief comments that give our reasons for this position.

1. One fundamental reason for including multiple-choice tasks in the accountability scores is to increase the scores' validity. Using enhanced multiple-choice tasks would allow the assessment of more higher thinking abilities within any given time frame. (Enhanced multiple-choice questions require the use of higher level thinking skills to process alternatives and answer correctly. Nonenhanced multiple-choice questions do not require higher level thinking skills to answer correctly). Further, using nonenhanced multiple-choice tasks would efficiently assess other desired knowledge spelled out by Kentucky's curricula (e.g., number concepts, Kentucky historical information, science principles and theories, and social studies generalizations).

2. Using enhanced multiple-choice tasks would assess students' abilities to apply and use concepts, principles, and problem-solving strategies at a certain level of cognition. These tasks would not be as amenable to "drill and practice" as nonenhanced multiple-choice tasks. Using them would broaden the curriculum areas the state assesses with little increase in time devoted to assessment. At the same time, the program could maintain high levels of cost-effective scoring and scorer reliability.

3. Discussions about eliminating the multiple-choice tasks have not recognized the use of enhanced multiple-choice tasks. These discussions seem to recognize the existence of only nonenhanced multiple-choice tasks (implying they require students to use only simple recall amenable to "drill and practice"). Further, the discussions have not provided a cogent argument for eliminating "drill and practice" as an instructional strategy for certain areas such as number concepts, Kentucky historical information, science principles and theories, and social studies generalizations.

4. The paper referred to above discusses reliability in a hypothetical way, focusing on internal consistency. This aspect of reliability is only part of the story,

however. Multiple-choice tasks are more reliable from the scoring perspective than are open-ended tasks. Also, since change or growth is a major interpretive framework in using the accountability index, reliability over time (sometimes called stability reliability) needs to be considered also. In addition, generalizability to the curriculum task domain is a significant reliability issue yet to be comprehensively studied by KDE.

5. Validity of assessment is the focal point when making policy decisions regarding what assessment modes to include. There is some validity evidence available that is unreported and some evidence that should be obtained before educators can judge the equivalence of two assessment modes and decide to eliminate one. First, for example, the correlations between the multiple-choice tests and the open-ended tests (data provided by Richard Hill) are not very high (approximately .66). This indicates that the two modes may be assessing related but different abilities. They are not interchangeable assessment modes as the above mentioned paper suggests. Second, the concept of matching tasks with topics listed in the content guidelines and in the academic expectations documents is a rather superficial approach to equivalence validation research. One would need to ascertain the various skills and abilities each mode assesses and then decide whether there was an opportunity to improve the tasks so they better assess the learning targets specified in the KDE outcomes and curriculum documents.

6. The paper referred to above concludes that the literature supports only a mixed picture of the impact of various formats on subgroups of students and that the multiple-choice tasks are "biased" in favor of males. These conclusions are unwarranted. The ACT study of the KIRIS performance assessment showed that KIRIS tasks favor males for science and mathematics, they favor females for reading, and they favor Caucasians over non-Caucasians. Perhaps the most relevant data for KIRIS on this issue is found in the NAEP national and state assessments. NAEP staff members have reported on differential performance at the task level and at the score levels for several subject matters and over several years. We suggest that this literature be reviewed very carefully before concluding there is mixed evidence. We suggest contacting the NAEP research director at Educational Testing Service (ETS), Princeton, NJ. The researchers at ETS have extensive experience in studying the differential performance of various student subgroups at different age levels on both the open-ended and the multiple-choice portions of the NAEP.

7. A broad and integrated approach to providing validity evidence should be presented to the State Board for Elementary and Secondary Education. The broad approach presents a comprehensive picture of the strengths and weaknesses of the inferences drawn from the accountability scores when various assessment modes are used. Armed with a more complete picture, the State Board for Elementary

and Secondary Education can make the necessary policy decisions about the technical merits of the current assessment mode configuration.

8. A disturbing part of the paper mentioned above is the suggestion that a future assessment strategy is to reduce the number of extended open-ended tasks and to increase the number of short-answer open-ended tasks. This would appear to reduce the validity of the Kentucky assessment scheme and to encourage more "drill and practice" in the long run. Schools already report that teachers are providing students with special training on how to answer short-answer questions. The latter may come about because the short-answer questions will focus on more specific facts and "bits" of information that teachers will soon learn to expect to appear on the assessment. They will soon develop strategies to drill students on these short-answer tasks. In addition, short-answer open-ended questions may in the long run be little more than multiple-choice questions without the "choices" or "distracters." This seems to be a regressive kind of assessment strategy.

9. We suggest that KDE study the feasibility of including more performance tasks and longer open-ended tasks in the assessment, while including a substantial number of multiple-choice tasks in the index as well. Such a strategy would appear to be in the spirit of the reform movement. Short-answer questions are not performance tasks. A multiple assessment strategy would likely increase the validity of the accountability index as a measure of the students' achievement in the schools and as a measure of the changing behavior of teachers in classrooms. It would better address, in our view, the "perceived message" issue than does narrowing the assessment modality. We think it would also mitigate against teaching to a narrow domain and perhaps cause teachers to vary their instruction desirably to help students demonstrate competence in a variety of ways.

Summary and Conclusions

Overall, KDE has addressed the legislative mandate to develop a high stakes performance assessment system. It has worked hard and demonstrated a high level of professionalism in developing KIRIS. The emphasis on performance assessment has improved the development of students' writing ability. The KIRIS assessment has also engaged students in the use of performance events.

In one piece of legislation, the Kentucky legislature tried to fundamentally revamp the education of children. With the passage of the KERA legislation, funding for education increased by 16-22 percent in Kentucky. The additional funds were designed, in part, to equalize funding across the school districts, as mandated by the Kentucky State Supreme Court.

However, KERA went far beyond the funding requirements of the court mandate. The state legislature set up a system of rewards and sanctions at the school level and mandated a

performance-based assessment. With the actions of the state legislature, Kentucky became the national leader in using performance-based assessments.

With the help of an outside contractor, ASME, the KDE was able to implement a performance-based assessment for all 4th, 8th and 12th grade students in the 1991-1992 school year. KERA allowed for a transitional testing period, to give KDE time to set up a performance-based assessment system. KDE and ASME **could** have implemented the legislative mandate for performance-based exams more slowly than they did. With such a quick implementation, the KDE and ASME moved Kentucky's assessment from multiple-choice to performance-based in one school year. This is a key achievement, but it brought about some special problems.

The system was implemented so quickly that it is not clear that stakeholders took the assessment system seriously during the baseline year. Student scores on the baseline assessment (1991-92) may have been **lower** than they should have been due to lack of familiarity of students and teachers with the test. Low student scores on the initial assessment may have made it easier for schools to reach rewards in the first biennium. Also, it appears that KDE was continuing to develop the KIRIS assessment using the experience it had gained and that new designs were emerging.

Surveys of constituent groups (Wilkerson & Associates, Inc., 1994; KIER, 1994) and what we learned through focus groups reveal that many stakeholders remain skeptical about KIRIS. We recommend that KDE work with teachers in order to improve teacher acceptance of the Kentucky reform movement. Right now, many teachers feel alienated from key decision making of the reforms. An alienated teacher work force is not going to help Kentucky students achieve at high levels.

Our Main Findings

For each of the study topics, we list points of both strengths and weaknesses. In the ensuing section we offer our ideas about what steps could be taken to strengthen the assessment program.

Consistency with Legislative Mandate

1. Overall, KIRIS is consistent with the Kentucky Educational Reform Act.

   1.1 On the major issue of performance-based, high stakes assessment, the Kentucky Department of Education has pursued the intent of the legislation. The Department was required to produce a fundamentally different kind of assessment for Kentucky students than the previously used state assessment tests. With KIRIS assessment, the Department of Education produced an assessment broadly consistent with legislative mandates.

   1.2 The legislation stipulated that the assessments were to provide the state with national comparisons similar to those provided by the National Assessment of Educational

Progress (NAEP--a federal assessment program providing benchmark information on student achievement). KDE provided national comparisons for two subject-matter areas in the 1992-1993 technical report. We understand that KDE plans to issue additional comparisons of KIRIS results with NAEP results when future NAEP results become available.

Understanding and Confidence of Stakeholders in the KIRIS Assessment

2. Most of the people who provided data for our study have some understanding of the rewards and sanctions component of the KIRIS assessment. However, specifics regarding rewards and sanctions are probably known only to a limited number of people (Department of Education personnel, superintendents and district assessment coordinators in some districts, some teachers, some principals, state legislators sitting on accountability committees, and some testing experts).

3. All the reviewed evidence suggests that principals, coordinators, superintendents, teachers, school council parents, public school parents, legislators, and the general public have serious questions concerning the legitimacy, validity, reliability, fairness, and usefulness of the KIRIS assessment. The groups surveyed perceived student performance on the KIRIS assessment as the measure **least** likely to provide a reliable indicator of student learning, compared to other commonly available indicators such as high school completion rate. The KDE will need to convince Kentucky educators that KIRIS is a sound basis for judging school effectiveness if this system is to become a valued part of the education reform process.

Involvement of Teachers and Principals in Design and Development of KIRIS

4. As described in the 1991-1992 Technical Report, advisory committees were established for reading, mathematics, science, and social studies. Representatives of KDE, teachers, curriculum coordinators, and Kentucky Education Association members sat on the committees. KDE added additional committees when other subjects were added to the assessment.

5. However, some of the teachers we communicated with were unaware of the input other Kentucky teachers had through these committees. Despite the committee system and the input of Kentucky educators into the review process of the KIRIS assessment, some teachers perceived that questions on the assessment were constructed by outsiders with little or no knowledge of Kentucky. Clearly, the perception of some teachers is at odds with the fact of educator involvement in KIRIS. This underscores the importance of continuing to involve and inform teachers and other educators in the ongoing process of assessment development.

Accuracy, Accessibility, and Clarity of Documentation

6. KDE has developed substantial technical information about KIRIS, given the early stage of development. As the program develops, there will be a continuing and growing need for technical information. We have outlined our view of what will be needed in our full report.

7. Also, there is a need for much better organization and improved balance of the information. While there is a considerable amount of technical data on the KIRIS assessment available in various places, it is difficult for anyone reviewing the program to compile all the relevant information. The technical reports do not provide a complete perspective on the weaknesses as well as the strengths of the KIRIS assessment results and on the accountability index.

Impact of the KIRIS Accountability Policies on Students, Teachers, and Schools

8. Students experienced more writing and group work under the reforms. Teachers, district assessment coordinators, and superintendents report almost unanimously that writing has improved, and the writing improvement was over and above what would have been expected of most school children of the same age.

9. Portfolios of students' written work have great instructional potential. However, portfolio scores vary considerably depending on which teacher is scoring the portfolio, making these scores less reliable than other forms of assessment.

10. The time and effort KDE invests in training teachers and that teachers spend marking the KIRIS portfolios, in our judgment, is probably useful as inservice education for teachers. We judge that the amount of instructional time teachers spend on the KIRIS assessment is reasonable.

11. The accountability index is influenced by factors beyond a school's control, but these are not taken into account when the index is interpreted. (Perhaps this is because the legislation does not require these factors to be taken into account.) Among the factors not considered are adequacy of resources, changes in the economic climate of a community, and changes in student mobility. However, the state maintains a mechanism by which a school's authorities can appeal such matters. That is, if a school believes the state's finding that the school failed to achieve the goal set for it by the state is due to factors beyond the control of the school, it can appeal the state's determination.

12. The accountability index does not provide teachers with timely feedback that is directly usable for improving classroom activities. While the index is not designed to provide such feedback, many of the educators with whom we communicated want more such feedback than the accountability index and the other aspects of KIRIS provide.

13. There is disagreement on the question of whether the system of rewards and sanctions will help improve the quality of education in Kentucky. District assessment coordinators think that rewards and sanctions will help improve education. Results for superintendents vary by survey. Teachers surveyed by KIER say the rewards and sanctions will **not** help improve education.

14. There is concern but as yet limited evidence about whether the administration of rewards and sanctions is fair to schools with large numbers of economically disadvantaged students, high turnover rates, or a very small number of students. We understand that KDE plans to provide further information on this important question in the near future.

15. The legislative intent of integrating assessment and feedback into the instructional process at every grade level has not been achieved. Teachers need more assistance than the Department of Education has so far been able to provide to embed performance assessments into the instructional process as was envisaged in the legislation.

Technical Adequacy of the KIRIS Assessment

16. On the whole we judged the KIRIS assessment tasks to be technically well crafted (the questions are clear and appropriate for the age group, the scoring rules are valid, and instructions are easy for students to follow).

17. The open-ended questions (those requiring a written answer) generally meet currently accepted technical item-writing standards for open-ended questions.

18. The district assessment coordinators and the superintendents overwhelmingly prefer a longitudinal approach (tracking the same group of individual students as they progress through the grades) over a cohort approach (comparing each group of 4th graders to those of previous years) for assessing a school's growth or change. In the opinion of the research team, longitudinal analysis gives a better picture of what impact the school is having on a group of students as it goes through the educational system, although it is more difficult and costly to implement. We believe that effective use of the longitudinal approach would require that assessments be administered at least to students at every other grade level and preferably at every grade level. It may also entail developing a growth scale on which a school's progress may be assessed. We note here that KDE made a deliberate policy decision early in the reform movement not to use the longitudinal model to evaluate growth in assessment scores.

19. The 1993-1994 KIRIS assessment included several modes: extended answer open-ended tasks, shorter answer open-ended tasks, multiple-choice tasks, portfolios, and performance events. This diversity of approaches is a strength of the scheme. The multiple modes of assessment approach is supported by educational assessment specialists because it enhances the validity of the results. Validity is enhanced by allowing students opportunities to demonstrate their abilities in a variety of ways over an appropriate range

of knowledge and skills. The proposed 1994-1995 KIRIS assessment will allow students fewer opportunities to demonstrate their abilities compared to the 1993-1994 KIRIS assessment due to the elimination of multiple-choice items. We think it was a mistake that KDE did not count the performances on multiple-choice items in computing the school accountability index; we think it would be a further mistake if KDE were to eliminate the multiple-choice items altogether. In addition, plans to increase the number of short-answer questions, instead of increasing the more in-depth performance components, narrows the modes of the assessment. The general point is that it will be desirable to broaden the assessment modes used.

20. The reliability of the accountability index is problematic for us. KDE has reported impressive reliabilities that reach or exceed .90, a level generally considered to be acceptable for use in high stakes decisions. However, because of the particular statistical model employed, there are unresolved questions about whether the high reliability estimates are indicative of the actual reliability. These concern, for example, whether to treat items or students as fixed, how agreements among raters are taken into account, and whether student scores should be estimated with regression.

21. Setting aside the issue of the statistical model for estimating reliability, it is clear that taken by themselves two of the three components of the KIRIS accountability index are not sufficiently reliable to be used in a high stakes assessment. These two components are the writing portfolio and performance events. We question whether the combination of these two components and the open-ended questions, which do evidence good reliability, give the Commonwealth a sufficiently reliable index for administering rewards and sanctions to schools. More reliability evidence is needed on this matter If the index is unreliable, then its validity is open to question since validity depends in part on reliability. The issues of the validity as well as the stability of the index require careful study, so that all stakeholders can be reassured that it provides a credible basis for administering rewards and sanctions or so that it can be corrected as needed.

Main Recommendations

The preceding assessment of strengths and weaknesses denotes that efforts to improve the KIRIS need to be continued if it is to provide a defensible basis for high stakes decisions and if it is to contribute productively to improving classroom instruction. In this section we offer our ideas about some of the steps that could be taken to address the continuing needs for improvement. While we have not had the time and resources to thoroughly develop these recommendations and to compare them to other possible improvement steps, we offer them in the spirit of helping Kentucky stakeholders to consider how best to continue improving KIRIS.

Additional Information and/or Reporting is Needed

1. There is a need to evaluate and address as appropriate concerns about the use of the accountability system. Among the concerns heard in our exchanges with Kentucky educators are that the current KIRIS assessment

   - narrows the curriculum
   - produces undue stress, especially on 4th grade teachers
   - yields an unstable index and unfair basis for accountability in those schools where individual student populations may vary widely from year-to-year and grade-to-grade
   - does not provide parents with reliable individual level student scores

2. The Commonwealth should investigate and report whether inner-city urban schools are being unfairly sanctioned because they have a more difficult educational task than the more stable schools. We understand that KDE plans to undertake such investigation following the completion of the first accountability cycle. However, this does not mitigate the fact that KIRIS results are being used in high stakes decision making before the needed evidence on the validity of KIRIS for this purpose could be obtained.

3. An index should be developed to report on the progress of students in meeting each of the four reform goals. It would also be desirable to report performance of schools on clusters of academic expectations.

4. Document and fully publicize the degree of interpretive and consequential validity of KIRIS. Also, document its instructional utility. Publicized reports should explain the appropriate cautions in using KIRIS results to claim educational improvement in Kentucky.

5. Continue to develop methods for reporting to schools on how they could use the KIRIS results to alter teaching and to improve student learning.

Training of Stakeholders

6. Given the weight of the writing portfolio in the accountability index, we recommend that the state continue to place great importance on the training of teachers to understand the deeper meaning of student writing and to score the writing portfolio.

7. Because of the instructional value of portfolios and the importance of having teachers seriously evaluate the best work of their students, teachers in Kentucky should continue to score the portfolios, even though scores of the same portfolios may vary from one teacher to the next and are, therefore, less reliable.

8. Expand on the steps being taken to involve and inform Kentucky educators about issues and developments in KIRIS. As much as possible, bring them into the partnership for developing and using a sound accountability index and helping to communicate KIRIS results to parents and other interested groups. Involve all Kentucky teachers in the process of crafting tasks that will be used in the operational assessment instruments.

9. Expand activities to help Kentucky teachers to incorporate the performance tasks and higher quality continuous assessments into their regular classroom instruction for all grade levels, as envisioned by the KERA.

Technical Issues

10. The technical reports should be organized so that an outside technical reader can evaluate the reliability and validity of the KIRIS results for achieving the uses and interpretations claimed for them. They also should summarize all the research results underpinning the program. There should be sections in the technical reports that point out problems and inconsistencies with the assessment. In general, they should include all the relevant technical information specified in the Standards for Educational and Psychological Testing (1985).

11. Beyond the requirements of the current standards, we suggest that KDE calculate and report reliability estimates for the accountability index based on a model that considers both students and items to be random sources of error, along with the estimates they now report using a model that considers students and items to be fixed factors. While this dual reporting would not resolve the debate about which model is the more appropriate, it would show readers the consequences to the reliability estimate of using one model or the other. We continue to believe that students and items should be considered as random sources of error in the generalizability model employed, since scores from one set of students and one set of assessment items are obtained in one year to set a threshold for evaluating the performance of a different set of students on a different set of assessment items in a subsequent year.

12. Continue to use the performance events. If the necessary approval can be obtained, we think it would be desirable to use the performance events to assess individual abilities to work collaboratively in groups as well as perform important learning targets as derived from the Kentucky goals and academic expectations. We note that KDE's current practice of not assessing students' ability to work effectively in groups is consistent with what the legislation permits.

13. Increase the priority and human energy resources devoted to analyzing data that support the technical underpinnings of the assessment results. This may or may not require an expanded staff. (We perceive that this change may already be under way, e.g., through studies relating the KIRIS results to American College Test [ACT] scores and through the Office for Education Accountability's study of the assessment.)

14. Provide evidence to demonstrate that the accountability index has a level of validity sufficient for use in high stakes decisions such as those affecting rewards and added resources such as planning grants and assignment of distinguished educators. Alternatively, if the necessary level of validity is not attained, do not continue to use the index for such decisions and actions until it is improved.

15. A key step toward improving validity will be to obtain external confirmation, as, for example, from the ACT, that the accountability index does manifest an acceptable level of reliability. Reliability is a necessary, but not sufficient, condition for validity. We recommend that KDE consult with a nationally recognized psychometrician who specializes in generalizability theory. The specialist should evaluate the statistical model, the estimated score procedures, and the design of the generalizability studies.

16. We think the decision not to include enhanced multiple-choice items in the index also limits the validity that could be attained, e.g., through improving both content coverage and reliability. We recommend, therefore, that KDE reassess the decision not to use enhanced multiple-choice test items, along with the short answer and performance assessments, in assessing student progress and computing the accountability index.

17. In the spirit of KERA's concern for authentic assessment, we also recommend that KDE at least consider increasing the emphasis on performance assessments that require speaking, developing products, organizing and planning activities, etc., compared to the heavy emphasis now given to performance assessments that require only written responses. We acknowledge that KDE and its contractor would need to conduct relevant research and development to fulfill this recommendation.

18. Consider using a longitudinal model to assess change in a school's accountability index.

Acknowledgements

We wish to thank the Kentucky Institute for Education Research, especially Roger Pankratz and Nila Weddle, for all of the help they have given to the research team over the last seven months. Without their help, this report would have suffered severe shortcomings. With their help, we have been in a better position to avoid factual and interpretive errors.

We also want to thank the hundreds of other people throughout the Commonwealth of Kentucky who played a role in this study. Unfortunately, only a few of these people can be mentioned in a report like this one.

During the qualitative data-gathering role, Robert Rodosky (Louisville), Duane Miller (Owensboro), Evonne Slusher (Bell County), and Joel Brown (Bowling Green) put together an impressive group of participants for the focus groups under a very tight timetable.

Many thanks to the Kentucky Association of Assessment Coordinators (KAAC) for allowing Fenster to "invite himself" to the May 20, 1994, meeting of the group. Because of that meeting, we added to the basic methodology of the study and decided to send surveys directly to stakeholder groups. If Fenster had not been able to attend that meeting, the idea for the survey would not have materialized. The surveys improved the quantity and quality of evidence presented in this report.

A special note of thanks to the 113 DACs and 70 superintendents who took the time from their busy schedules to answer an intensive questionnaire about their experiences with KERA and KIRIS. Without the time and effort of these people, the study would have been significantly weaker.

We thank Edward Reidy of KDE and Richard Hill and Amy Sosman of ASME for taking the time from their busy schedules to provide documents and to repeatedly answer our telephone questions on the KIRIS assessment.

We also recognize the long and hard work put into the KIRIS assessment system by ASME and KDE. Performance assessments are not commonly used. The problems with these new kinds of assessments have not yet been worked out technically nor operationally. It would have been easy for ASME and KDE to go slowly when implementing a new performance assessment system. ASME and KDE took the tougher road, bypassed the transitional testing period, and implemented the legislatively mandated performance-based system immediately.

The Evaluation Center has been pleased to play a role in the first overview evaluation of the KIRIS assessment. The Center has conducted many program evaluations of statewide assessments over the last 20 years. The KIRIS approach to educational accountability is interesting and useful. A great deal is riding on its quality and impact. We wish KDE and Kentucky's school personnel well as they continue the task of developing a sound system of assessment to help drive and document educational reforms in the Commonwealth.

REFERENCES

American College Testing Research Division. (1994). A study of core course-taking patterns for Kentucky ACT-tested graduates of 1991-1993 and an investigation of the relationship-based assessment results and ACT-tested Kentucky graduates of 1992. Iowa, City, IA: Author.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). Standards for educational and psychological testing. Washington, DC: American Psychological Association.

Coe, P., Leopold, G., Simon, K., & Williams, J. (1994). Perceptions of school changes: Interviews with Kentucky students. A Report Submitted to the Kentucky Caucus of the Appalachia Educational Laboratory Board of Directors. Charleston, WV: AEL.

Guskey, T. R. (1994). High stakes performance assessment: Perspectives on Kentucky's educational reform. Thousand Oaks, CA; Corwin Press.

Horizon Research International. (1994). A survey of legislators on Kentucky instructional results information system (KIRIS). Legislative Research Commission, Office of Education Accountability. Frankfort KY: Author

Hughes, K,R., & Craig, J.R. (1994, November). Using performance assessment achievement data to evaluate a primary instructional program. Paper presented at the annual meeting of the American Evaluation Association, Boston, MA.

Kentucky Department of Education. (1994). Kentucky instructional results information system: 1992-93 technical report. Frankfort, KY: Author.

Kentucky Department of Education. (1993). Kentucky instructional results information system: 1991-92 technical report. Frankfort, KY: Author.

Kingston, N. (1994). Further considerations of issues related to the inclusion of multiple-choice items in KIRIS reported scores. Frankfort, KY: Kentucky Department of Education.

Koretz, D., Stecher, B., Klein, S., & McCaffrey, S. (1994). The Vermont portfolio assessment program: findings and implications. Educational Measurement: Issues and Practices, 13(4), pp. 5-16.

Linn, R. L. (1994). Performance assessment: Policy promises and technical measurement standards. Educational Researcher, 23(9), pp. 4-14.

Linn, R.L., Baker, E.L., & Dunbar, S.B. (1991). Complex, performance-based assessment: expectations and validation criteria. Educational Researcher, 20(8), pp. 15-21.

Meyer, R. (1994). Educational performance indicators: A critique. Unpublished paper. Harris Graduate School of Public Policy Studies, The University of Chicago.

The Kentucky Institute for Education Research. (1994). An evaluation of the progress of KERA: The judgments, opinions and perspectives of Kentucky school superintendents. Frankfort, KY: Author.

North Central Regional Education Laboratory. (1995). Choosing persons who are likely to succeed and remain as teachers in urban schools serving children and youth in poverty: A proposal to examine and extend the technique, adequacy and utility of the Haberman technique. Oak Brook, IL: Author.

Rose v. Council for Better Education, Inc. KY 88-SC-804-TG (September 28, 1989).

Wilkerson & Associates, Ltd. (1994). Statewide education reform survey. Louisville, KY: Author.

APPENDIX A

Evaluation Procedures and Sources of Evidence

In the course of conducting this evaluation, we

Responded to a Request for Proposals from the Kentucky Institute for Education Research. March 21, 1994. (Fenster)

Revised the original Request for Proposals to address concerns for KIER. April 6-9, 1994. (Stufflebeam, Fenster)

Met with a Technical Assessment Group in the KIER offices (Skip Kifer, Ben Oldham, and Ella Simmons). April 27, 1994. (Stufflebeam, Fenster)

Met with Sharon Solomon, President of Kentucky PTA. April 27, 1994. (Stufflebeam, Fenster)

Met with Sharon Felte Comer, Kentucky Education Association. April 27, 1994. (Stufflebeam, Fenster)

Attended focus group meeting of Jefferson County teachers and principals. April 28, 1994. (Stufflebeam, Fenster)

Met with Robert Rodosky, Assessment Coordinator of Jefferson County. April 28, 1994. (Stufflebeam, Fenster)

Met with Christy Maloney, ASME, in Louisville. April 28, 1994. (Stufflebeam, Fenster)

Received contract from KIER. April 28, 1994. (Stufflebeam)

Met with Ken Scott, Kentucky School Boards Association. April 28, 1994. (Stufflebeam, Fenster)

Met with Ed Reidy and Neal Kingston. April 28, 1994. (Stufflebeam, Fenster)

Met with Ken Draut, Assessment Coordinator for Henry County. April 28, 1994. (Stufflebeam, Fenster)

Met with Penny Sanders, Kentucky Office of Education Accountability. April 28, 1994. (Stufflebeam, Fenster)

Witnessed a KIRIS performance event assessment in a suburban school in Fayette County. April 29, 1994. (Stufflebeam, Fenster)

Studied the KIRIS background materials. May-December 1994. (Stufflebeam, Fenster, Nitko, Meyer, Wiersma)

Conducted document review of the KIRIS and KERA materials. May-December 1994. (Stufflebeam, Fenster, Nitko, Meyer, Wiersma)

Interviewed Gerald Hutchins, Assessment Coordinator, Fayette County. May 12, 1994. (Fenster)

Interviewed Donna Shedd, PPIE, in Louisville. May 13, 1994. (Fenster)

Discussed progress of study with Roger Pankratz, KIER. May 13, 1994. (Fenster)

Interviewed Leon Mooneyhan, Superintendent of Shelby County. May 13, 1994. (Fenster)

Interviewed Jackita Neill, Assessment Coordinator, Henderson County. May 16, 1994. (Fenster)

Interviewed Jack Rose and Joy Waldrop, Superintendent and Assessment Coordinator, Calloway County Schools. May 17, 1994. (Fenster)

Interviewed Don Sparks and Lenna Austin, Superintendent and Assessment Coordinator of Mayfield Independent Schools, respectively. May 17, 1994. (Fenster)

Conducted focus group session held in Owensboro. Participants included assessment coordinators, principals, and teachers. May 18, 1994. (Fenster)

Interviewed Vicky Clemens, Assessment Coordinator and Instructional Supervisor, Calloway County Schools. May 19, 1994. (Fenster)

Attended and made brief presentation describing study to meeting of Instructional Supervisors of Northern Kentucky. May 19, 1994. (Fenster)

Attended KAAC meeting in Elizabethtown. The idea for a survey to the DACs came from this meeting. First draft of DAC Survey written during this meeting. May 20, 1994. (Fenster)

Interviewed Kirby Wright, Instructional Supervisor, Mason County. May 23, 1994. (Fenster)

Meeting in Pike County cancelled. Offices closed due to state primary. May 24, 1994.

Conducted focus group in Bell County. Participants included assessment coordinators, superintendents, principals, teachers, members of the Pritchard Committee for Academic Excellence, and the general public from six districts in southeastern Kentucky. May 25, 1994. (Fenster)

Conducted focus group at Greenwood High School, Bowling Green. Participants included Pat Gutherie, Assessment Coordinator of Warren County; the principal of Greenwood High School; and 3 English teachers. May 26, 1994. (Fenster)

Conducted focus group in Bowling Green Independent School District. Participants included Superintendent Joel Brown, the financial officer, principals, and teachers. May 26, 1994. (Fenster)

Interview with Joe Hignite, Instructional Supervisor, Perry County. May 27, 1994. (Fenster)

Drafted versions 2 thru 11* of District Assessment Coordinator Survey. June 1994. Received input from Judy Tabor, Neal Kingston, and Brian Gong of KDE on first and fourth draft of survey. Received feedback from participants of KAAC meeting on first draft of survey. Received feedback from Roger Pankratz on first, fourth, and tenth versions of survey. (Stufflebeam, Fenster, Wiersma)

Mailed District Assessment Coordinator Survey. June 20, 1994. (Fenster)

Mailed Superintendent Survey. June 27, 1994. (Fenster)

Prepared questions for trip to Dover, NH. July 6, 1994. (Fenster, Wiersma)

Interviewed Amy Sosman and Richard Kahl of ASME, Dover, NH. July 14, 1994. (Nitko, Meyer, Fenster)

Interviewed Richard Hill, President of ASME. July 15, 1994. (Nitko, Meyer, Fenster)

Developed survey for PTA Leadership Conference. July 18-19, 1994. (Fenster, Nitko, Meyer, Wiersma)

Attended PTA Leadership Conference. Louisville. July 22-23, 1994. (Fenster)

Interviewed Jack Foster by telephone from Kalamazoo, MI. August 1994. (Fenster)

Analyzed the DAC Survey. August 8, 1994. (Fenster)

Analyzed Superintendent Survey. August 9, 1994. (Fenster)

Prepared written analysis of Superintendent and DAC surveys. August 1994. (Wiersma)

Received input from Wiersma on first draft of report. August 22, 1994.

Received input from Nitko on first draft of report. August 23, 1994.

Responded to Neal Kingston's paper. August 25, 1994. (Nitko)

Sent first draft of report to Roger Pankratz. August 26, 1994. (Fenster).

Analyzed data from PTA survey. September 8, 1994. (Fenster)

Prepared written analysis of PTA survey. September 12-14, 1994. (Wiersma).

Received clarification on aspects of the KIRIS assessment from Neal Kingston and Scott Trimble. September 14 and 16, 1994. (Fenster)

Held a conference call to receive input from Nitko on second draft of report. September 15, 1994. (Fenster)

Drafted second draft of report. September 9-19, 1994. (Stufflebeam, Fenster)

Presented a second draft of final report to KIER Board. September 20, 1994. (Stufflebeam, Fenster)

Discussed report with Ray Nystrand, Dean of Education, University of Louisville. September 20, 1994. (Stufflebeam)

Received input from members of the evaluation team on the September 20, 1994 draft of report. October-November 1994.

Planned with Roger Pankratz about how to make this evaluation of KIRIS a constructive force for improving the utility and credibility of the assessment. November 1994. (Stufflebeam)

Completed third draft of final report to account for comments from the evaluation team. November 1994. (Fenster)

Received feedback from the KIER Board on the November draft of the report. December 1994. (Stufflebeam, Fenster)

Received feedback from Roger Pankratz on the November draft of the report. December 1994. (Stufflebeam, Fenster)

Received feedback from Kentucky reviewers on the November draft of the report. December 1994. (Stufflebeam, Fenster)

Completed fourth draft of final report to account for comments from the Kentucky reviewers. December 27, 1994. (Fenster)

Received feedback from KDE on the December 27, 1994, draft of the report. January 4, 1995. (Stufflebeam)

Received feedback from Roger Pankratz on the December 27, 1994 draft. January 1995. (Stufflebeam)

Received comments from Ray Nystrand on the December 27, 1994, draft. December 31, 1994. (Stufflebeam)

Received comments from Anthony Nitko on a December draft. January 5, 1995. (Fenster)

Revised report to account for reviewer comments. January 6-13, 1995. (Stufflebeam, Fenster)

Completed fifth draft of final report. January 13, 1995. (Stufflebeam, Fenster)

Held conference call to receive feedback on the fifth draft of final report. January 16, 1995. (Stufflebeam, Nitko, Fenster)

Received feedback from Skip Kifer on the fifth draft of the report. January 1995. (Stufflebeam, Nitko, Fenster)

Received feedback from Wiersma on the fifth draft of report. January 30, 1995. (Fenster)

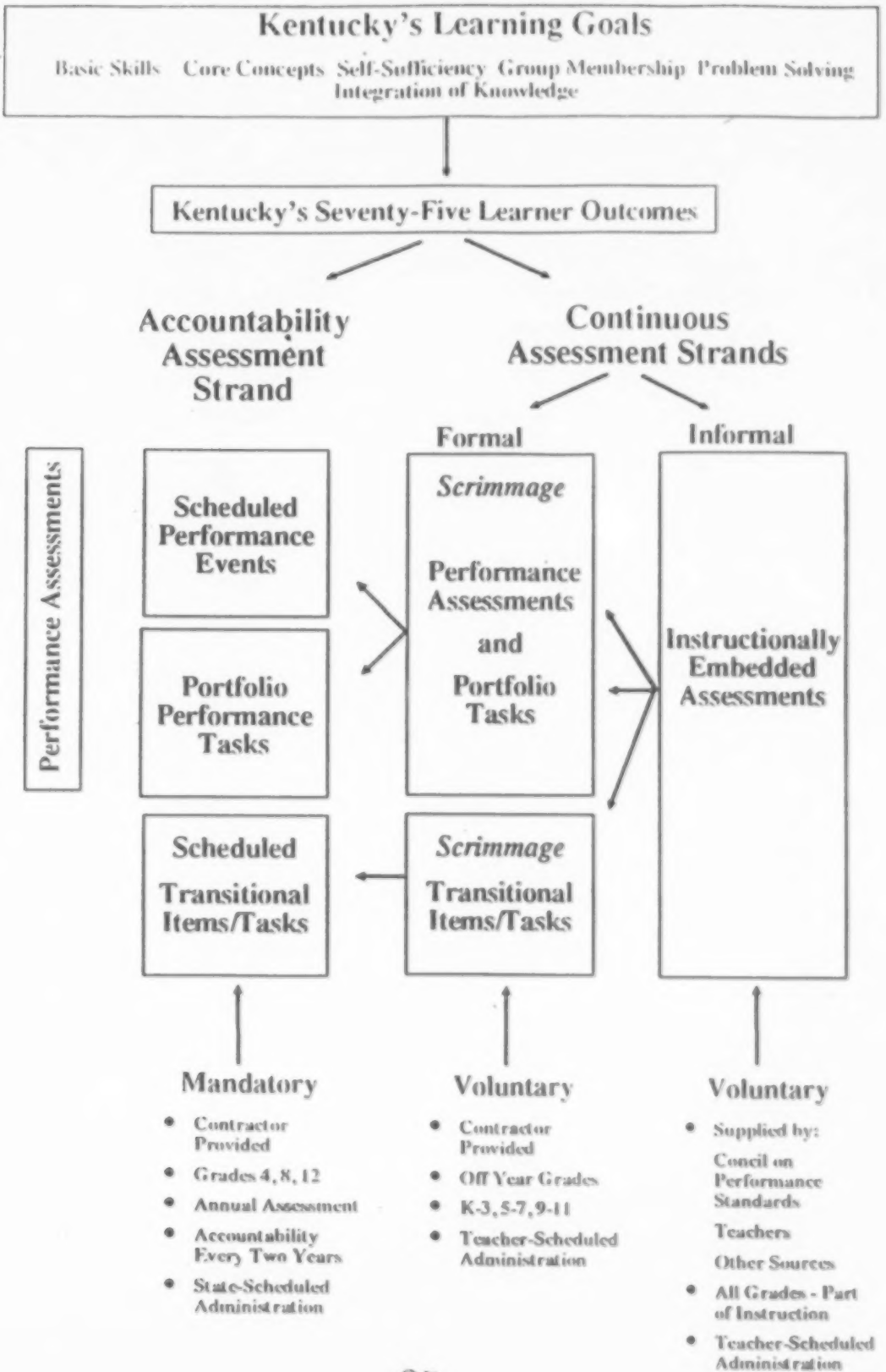
Received feedback from KDE on the fifth draft of the report. February 2, 1995. (Stufflebeam)

Received feedback from Nitko on the fifth draft of the report. February 4, 1995. (Fenster)

Redrafted report. February 4-5, 1995. (Stufflebeam, Fenster)

Sent final report to KIER. February 8, 1995. (Fenster)

APPENDIX B


Kentucky's Learning Goals
Basic Skills Core Concepts Self-Sufficiency Group Membership Problem Solving Integration of Knowledge
Kentucky's Seventy-Five Learner Outcomes
Accountability Assessment Strand
Continuous Assessment Strands
Formal
Informal
Performance Assessments
Scheduled Performance Events
Portfolio Performance Tasks
Scheduled Transitional Items/Tasks
Scrimmage
Performance Assessments and Portfolio Tasks
Scrimmage
Transitional Items/Tasks
Instructionally Embedded Assessments
Mandatory
Contractor Provided
Grades 4, 8, 12
Annual Assessment
Accountability Every Two Years
State-Scheduled Administration
Voluntary
Contractor Provided
Off Year Grades
K-3, 5-7, 9-11
Teacher-Scheduled Administration
Voluntary
Supplied by:
Concil on Performance Standards
Teachers
Other Sources
All Grades - Part of Instruction
Teacher-Scheduled Administration

APPENDIX C

KENTUCKY DEPARTMENT OF EDUCATION
CAPITAL PLAZA TOWER • 500 MERO STREET • FRANKFORT, KENTUCKY 40601
Thomas C. Boysen, Commissioner

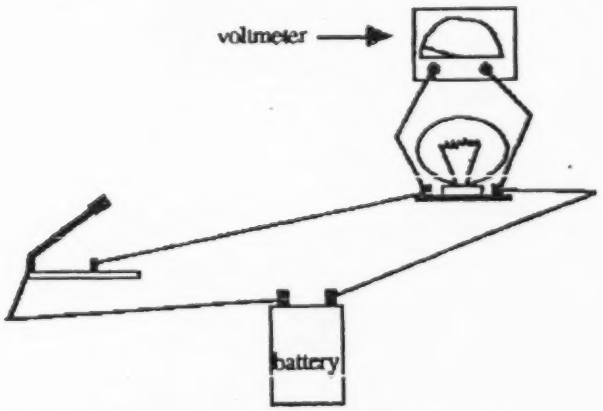
April 8, 1994

Mr. Richard Innes
2836 Deerfield Drive
Villa Hills, KY 41017-4470

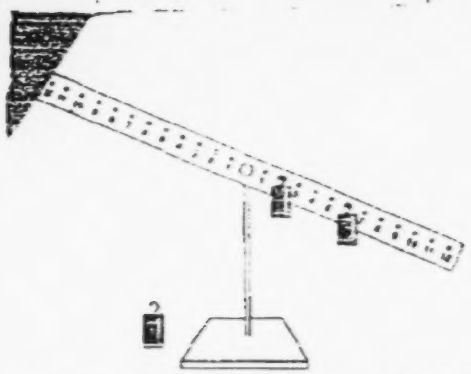
Dear Mr. Innes,

Thank you for your interest regarding the KIRIS Assessment Program and specifically, the test item you questioned. Since this assessment is unique in its design and focus, I'm sure you appreciate the difficulties involved in the construction process. Before I address the assessment item you questioned, I would like to apologize for the time it has taken to respond to your inquiry. Your concern for Kentucky students and the assessment is admirable and welcome and our response should have been more prompt.

The question, as written, is incorrect. In order to correct it, the voltmeter should have been connected across the component to be measured as shown in the figure below:



This configuration would, as I am sure you already know, allow the light to illuminate and the voltmeter to measure the voltage. Of course, if an ammeter was used instead of the voltmeter, there would have been no problem with the question.



Look at the picture shown above. There are two weights hanging on the right side of the scale. If only one weight were hung on the left side of the scale, where must it be hung to make the scale balance? (All three weights are of equal mass.)

A. 4
B. 8
C. 10
D. 12

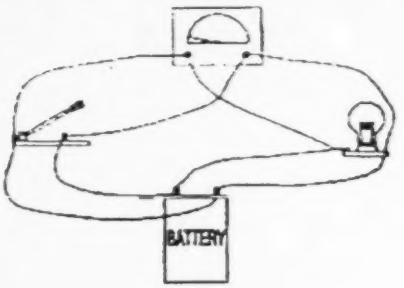
Weather reports often give the chance of rain for a given day. This figure is computed by determining

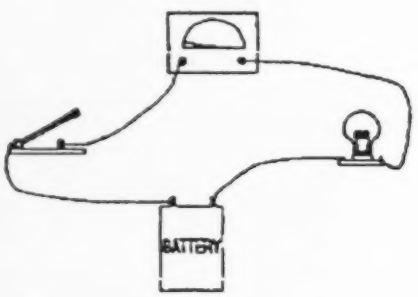
A. the extent to which the clouds are saturated.
B. how often it rained in the past when the conditions were similar.
C. the percent of the earth's surface that is likely to receive rain.
D. the portion of time during the day when rain is expected.

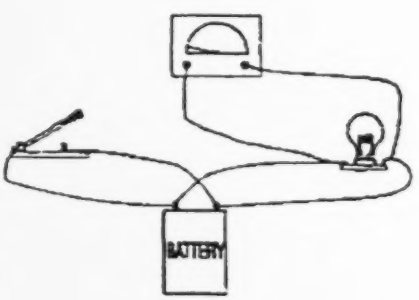
Burning fossil fuels may cause an increase in the atmospheric carbon dioxide. What effect might such an increase have on living things?

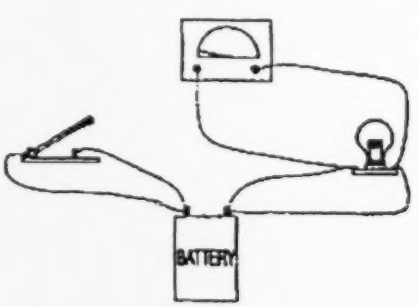
A. Animals would suffocate because the carbon dioxide would force oxygen out of the atmosphere.
B. Plants would die because they wouldn't get enough sunlight.
C. Some kinds of plants and animals would die because the climate would become warmer.
D. All life would die because carbon dioxide is poisonous to living things.

40. In which picture are the bulb, the battery, the switch and the voltmeter all connected properly?

A.


B.


C.


D.


While this question was written the first year of the Kentucky KIRIS assessment and I was not involved in its development, I am fairly certain that the intent was to assess the students' ability to recognize a complete circuit and not students' ability to correctly connect a voltmeter to measure voltage across a conductor. At any rate, the question was no longer used after the 1991-92 year and the multiple choice items have never been used for accountability purposes (i.e. - the students' grades and schools' scores were not affected in any way).

Regarding the appropriateness of the question, as a former physical science teacher (9th grade), my students have used both ammeters and voltmeters in lab activities. Experimenting with the meters and the ways to connect them helped my students understand a little bit more about current and voltage, as well as why the meters needed to be connected in a particular way. I do not agree that knowing how to measure voltage or current is too sophisticated for twelfth grade students. I think you will find measurement of current and voltage typical concepts for most physical science texts.

As to your inquiry about lay participation, a Content Advisory Committee (CAC) composed entirely of Kentucky teachers meets three times a year to construct, evaluate, and review assessment items for the KIRIS test. At this point, there are no provisions for lay participation in the development process. The CAC works very hard to insure that a fair, accurate and effective assessment is constructed for Kentucky students. The fact that few of the KIRIS assessment questions written over the past four years have come under fire is a testament to the quality of the test. When this group meets again in June, I promise to inquire as to the possibility of enlisting the assistance of additional technical experts to review assessment items.

I hope I have responded to the concerns mentioned in your letter. Please feel free to contact me or my Division Director, Tim Moore, if you have any additional questions.

Thank you,

Kevin Stinson
PRISM Science Consultant

c: Tim Moore